碩 士 學 位 論 文

다중 분할표 분석에서 SVS-BSR 방법을 이용한 고차원 상호작용 추론

高麗大學校 大學院

應用統計學科

嚴 智 閶

2020年 02月

全 秀 榮 敎 授 指 導

碩 士 學 位 論 文

다중 분할표 분석에서 SVS-BSR 방법을 이용한 고차원 상호작용 추론

이 論文을 統計學 碩士學位 論文으로
提出함.

2020年 02月

高 麗 大 學 校 大 學 院

應 用 統 計 學 科

嚴 智 閏 (印)

嚴　智　閏의 統計學 碩士學位論文

審査를 完了함.

2020年 02月

委員長 김두영 (印)

委　員 홍 승 만 (印)

委　員 임 성 수 (印)

요 약 문

최근 데이터의 양이 증가하면서 분할표의 변수의 수도 증가하여 변수들의 고차 상호작용효과에 관심이 많아졌다. 하지만 변수의 수가 관측치 수만큼 커짐에 따라 과적합의 문제가 발생하는 경우가 많다. 이를 극복하기 위해 본 연구는 변수 검사과정(sure variable screening, SVS)을 이용한 베이지안 부분집합 회귀(Bayesian subset regression, BSR) 방법인 SVS-BSR 방법을 제안한다.

제안된 SVS-BSR 방법에서 제시하는 최대 사후모형은 최소 확장 베이지안 정보기준(minimum extended Bayesian information criterion) 모형과 근사적으로 같다. 또한 BSR 사후분포로부터 효율적인 표본추출을 위해 확률적 근사 몬테카를로(stochastic approximation Monte Carlo) 알고리즘을 이용한다. 더 나아가서, 변수 검사 과정은 주변 포함 확률에 근거하여 제안된다. 유의하게 선택된 변수들 중 실제로는 유의하지 않으나 선택된 변수들의 기대 비율인 FDR로 선택된 변수들을 살펴본다. 용인할 수 있다고 정한 기준값보다 작은 거짓 발견율을 선택하여 선택된 변수들을 검사한다.

제안된 변수 검사과정을 이용한 SVS-BSR 방법의 우수성을 보기 위해, 네 개의 분할표 자료에 대해 SVS-BSR 방법을 기존의 벌점화우도 방법들인 능형(ridge), 라소(lasso), 엘라스틱넷(elastic net) 방법들과 비교를 한다. 비교 결과, 모든 예제에서 SVS-BSR에 의해 선택된 모형들이 변수의 수가 다른 방법에 비해 작으면서 가장 작은 제곱근 하나남기기 교차검증(root leave-one-out cross validation) 값을 가짐에 따라 SVS-BSR 방법이 벌점화우도 방법들보다 우수함을 알 수 있었다. 또한 벌점화우도 방법들은 분할표의 차원이 커질수록 결과가 좋지 않았다.

핵심어 : 다중 분할표, 고차 상호작용 효과, 베이지안 부분 집합 회귀, 확률적 근사 몬테 카를로 방법, 변수 검사과정, 변수 선택, 거짓 발견율

목 차

표 목 차

제 1 장 서 론

통상적인 일반 선형모형(general linear model)은 독립변수와 종속변수 사이의 선형성, 오차항의 정규성, 독립성과 등분산성 4가지를 기본 가정으로 한다. 하지만 실제 자료는 연속형이 아닌 경우가 많다. 이런 기본 가정들이 적용될 수 없는 경우를 위해 일반 선형모형을 확장한 일반화 선형모형(generalized linear model)이 다양한 분야에서 사용되고 있다. 일반적으로 변수의 수 P_n이 관측치 수 n보다 더 클 때 ($P_n > n$) 일반화 선형모형에서 고차원(high-dimension)의 문제가 있다고 보지만, 본 연구에서 다루는 분할표의 로그 선형모형에서처럼 변수의 수가 관측치 수만큼 큰 경우 ($P_n \approx n$)에도 고차원의 문제가 있다고 볼 수 있다.

최근 데이터의 양이 점점 증가하면서 분할표 변수의 수와 그 변수의 범주 또한 증가하여 변수들의 고차 상호작용 효과에 관심이 많아졌다. 하지만 변수의 수가 관측치 수만큼 커짐에 따라 과적합(overfitting) 등 고차원의 문제점이 발생하는 경우가 많다. 따라서 수많은 변수들 중 설명변수로 어떤 변수를 선택하느냐는 좋은 모형을 수립하는 데에 있어서 중요한 문제가 된다.

고차원 일반화 선형모형에서 특별한 관심 분야는 반응변수의 중요한 인과 특징(causal feature)을 형성하는 설명변수들의 부분집합을 찾는 변수선택 문제에 있다. 예를 들어, 질병 발생에 대한 유전적 관련성을 분석할 경우 유전자 간의 복잡한 교호작용과 유전자와 환경적 요인과의 교호작용을 고려하는데, 흔히 사용하는 통계분석 방법인 로지스틱 회귀분석에서는 SNP(single nucleotide polymorphism)의 숫자가 많아지면 그것들 사이의 복잡한 교호작용 효과에 대한 해석이 어려워진다(Foulkes, 2005). 고차원의 변수선택 문제는 일반적으로 벌점화우도(penalized likelihood)와 베이지안 방법으로 해결될 수 있다(Liang et al., 2013).

벌점화 우도 방법에는 능형회귀(ridge regression), 라소(lasso), 그리고 엘라

스틱넷(elastic net) 세 가지가 있다. 능형회귀는 관측치 수보다 고려해야 할 설명변수의 수가 더 많을 때 최소 제곱 추정량을 구하기 위해 Hoerl and Kennard(1970)에 의해 고안되었다. 능형회귀는 고려해야 할 설명변수들의 다중공선성이 높아 회귀계수의 과잉 추정이 우려될 때 회귀계수가 평균 쪽으로 줄어드는 효과를 가진다. 라소는 Tibshirani(1996)에 의해 고안된 방법으로 회귀계수에 벌점을 부과하여 계수를 축소하는 방법이라는 측면에서 능형회귀와 동일하지만, 모든 설명변수를 고려하는 능형회귀와 다르게 변수선택의 기능까지 있다는 점에서 큰 장점을 갖는다. 엘라스틱넷은 설명 변수의 수가 관측치를 넘어설 수 없도록 고안된 라소를 보완하여 라소와 능형회귀의 제약식을 동시에 고려하는 방법론이다.

고차원 문제를 해결하기 위해 여러 연구(Sun et al., 2007; Oh, Le, 2013; Lee, Kwon, 2015; Jo, Cheon, 2015)와 고급 몬테카를로 방법들(Hans et al., 2007; Bottolo, Richardson, 2010)이 제안되었다. 대부분의 베이지안 방법들은 부분집합 모형의 음의 로그 사후확률이 근사적으로 전통적인 부분집합 모형 선택 통계량(예로 Mallows' C_p, AIC, BIC)이 되고, 적절한 사전분포를 이용하여 구축한 베이지안 부분집합 회귀(Bayesian subset regression, BSR) 모형이 전통적인 베이지안 회귀모형보다 예측에 있어 더 효율성이 좋다는 연구가 있다(Liang et al., 2001; Hsu, 1995). Liang et al.(2013)은 고차원에서 small-n-large-p의 경우에 전통적인 방법의 문제점을 극복하고자 최소 EBIC(Extended Bayesian Information Criterion)와 동일한 최대사후모형인 새로운 BSR 모형을 제안하여 기존의 벌점화우도 방법들인 라소, 엘라스틱넷 등 보다 더 우수하다는 것을 보였다.

또한 Liang et al.(2013)은 BSR 방법 내에서 예측변수들의 주변 포함확률에 기반한 변수 검사과정인 다중 가설검정 기반 변수 검사과정(multiple test -based sure variable screening procedure)을 소개했다. 그들은 비록 부분집합 모형의 음의 로그사후확률이 근사적인 특징을 갖고 있다고는 하지만, 추정 값의 기준이 무엇이 될지에 대해서는 분명하지 않다는 점을 지적했다. 주변 포함

확률은 모든 예측변수들의 결합정보를 포함하기 때문에 다른 검사과정들보다 더 좋은 결과를 보인다. 또한 모든 변수들을 다시 살펴보고 검사하는 과정이 필요하다고 판단하여 거짓 발견율인 FDR(False discovery rate)을 이용한 변수 검사과정을 거친다.

분할표에서의 포아송 로그 선형모형은 반응변수에 상호작용 효과를 포함한 여러 범주의 설명변수들이 어떤 영향력을 미치는지에 대해 파악하는 통계모형으로 널리 사용된다. 실제 고차원 분할표에서 정확한 분석을 위해 고차의 상호작용 효과를 포함시킬 필요가 있다. 만약 관련 없는 상호작용 효과 변수를 포함한다면 로그선형모형은 잘못된 통계적 추론을 가져다 줄 수 있으므로 설명력이 높은 유의한 상호작용 효과를 정확하게 찾는 것은 매우 중요한 문제이다. 하지만 변수의 수 P_n이 관측치 수 n만큼 크거나($P_n \approx n$) 또는 n보다 더 클($P_n > n$)경우 자료를 과적합(overfitting)할 수 있는데, 과적합으로 추정된 모형은 일반적으로 영향력이 없는 변수를 다수 포함하게 되어 예측력이 좋지 않다. 이처럼 고차원 모형 분석을 위한 전통적인 통계방법은 종종 정확한 계산에 어려움이 있어 모형 추정에 문제를 가지고 있다.

기존 방법들은 일반화 선형모형에서 고차원인 경우, 정확하지 못한 추론 결과를 제공한다. 특히 고차원 분할표 분석을 위한 포아송 로그선형모형에서 반응변수들에 유의하게 작용하는 고차의 상호작용 효과를 찾기 위해 진행된 연구가 적다. 따라서 본 연구는 고차원 분할표 분석에서 유의한 고차 상호작용 효과를 찾기 위해 자기조절능력(self-adjusting ability)이 있는 확률적 근사 몬테카를로(stochastic approximation Monte Carlo, SAMC; Liang et al., 2007) 알고리즘과 베이지안 부분집합 회귀(Bayesian subset regression, BSR; Liang et al., 2013), 그리고 변수 검사과정(Sure Variable Screening, SVS; Liang et al., 2013)을 이용한 방법을 제안하고자 한다.

본 논문의 구성으로는 2장에서 일반화 선형모형의 조건부 추론에 대해 소개한다. 3장에서는 고차원 분할표 분석을 위한 기존 방법론들인 능형회귀, 라소, 엘라스틱넷, SAMC를 이용한 BSR 방법을 소개하고, 4장에서는 다중 가설검정

기반 변수 검사과정을 이용한 베이지안 부분집합 회귀(SVS-BSR)을 제안한다. 5장에서는 모의실험 자료를 바탕으로 제안한 변수선택 모형을 기존의 방법들과 비교한다. 6장에서는 고차원 분할표 분석에서 사용되는 다양한 크기의 실제 예제들을 적용해 보았다.

제 2 장 분할표 분석을 위한 포아송 로그선형모형

이 장에서는 분할표에 대한 조건부 추론을 실행하기 위해 필요한 표기법과 설정의 개요를 설명한다. $D^n = \left\{(x_1, x_2, \cdots, x_{P_n}) : x_i \in R^n\right\}$을 P_n개의 설명변수로 구성되어 있는 n개의 관측치 데이터 셋이라고 하자. 두 개 이상의 범주형 변수들에 따라 교차 분류된 도수표를 분할표라 한다. 본 연구는 고차원 분할표 분석을 위해 포아송 로그선형 모형(Poisson log-linear model)을 적용한다.

예로 IJK개의 셀로 구성되는 세 개의 범주 (X, Y, Z)에 대한 삼차원 분할표를 살펴보자. 각 셀 값이 $\{X_{ijk} = x_{ijk},\ i = 1, \cdots I,\ j = 1, \cdots, J,\ k = 1, \cdots K\}$인 분할표 X를 사용한다. X의 포아송 분포는 일반적으로 아래와 같이 표현된다.

$$f(x|\mu) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{e_{ijk}^{-\mu} \mu_{ijk}^{x_{ijk}}}{x_{ijk}!}. \quad \cdots(1)$$

여기서 μ_{ijk}는 셀(i,j,k)의 기대 빈도수이고, $\mu = (\mu_{ijk})$, $x = (x_{ijk})$이다. 로그선형모형은 다음과 같이 로그선형함수에서 기대 셀 빈도수를 모수화한 포화모형(saturated model)이다.

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}. \quad \cdots(2)$$

여기에서 $\lambda_{i,}$ $\lambda_{j,}$ λ_k를 각각 행(row), 열(column), 층(layer) 효과, λ_{ij}, $\lambda_{jk,}$ λ_{ik}를 1차 상호작용 효과, λ_{ijk}를 2차 상호작용 효과라 한다. 더미 변수들에 대하여, λ_{ijk}^{XYZ}는 X에 대한 i번째 더미 변수, Y에 대한 j번째 더미 변수, 그리고 Z에 대한 k번째 더미 변수의 곱의 계수이다.

식(1)과 (2)에 의해 정의된 분할표에서 $IJK(=n)$ 관측치를 가지고 있는 로그선형모형의 중복되지 않은 모수들의 전체 수는 아래와 같고, 이것은 셀의 전체

숫자 총 수이다.

$$1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1) \quad \cdots(3)$$
$$+(J-1)(K-1)+(I-1)(J-1)(K-1)=IJK.$$

이 모형은 관측치 수만큼 많은 모수들을 가지고 있고 이것은 포화모형이다. 식(2)에서 어떤 모수들을 0이 되게 설정하면 이전에 소개된 모형들이 된다. 아래의 표는 이 모형들 중 몇 개를 열거한 것이다. 모형을 나타내기 쉽도록 표1은 각 모형의 각 변수에 대해 가장 높은 순서의 항을 열거하였다.

<표 1> 삼차원 분할표에서 로그선형모형

Loglinear Model	Symbol
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z$	(X,Y,Z)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}$	(XY,Z)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{jk}^{YZ}$	(XY,YZ)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{jk}^{YZ}+\lambda_{ik}^{XZ}$	(XY,YZ,XZ)
$\log\mu_{ijk}=\lambda+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{jk}^{YZ}+\lambda_{ik}^{XZ}+\lambda_{ijk}^{XYZ}$	(XYZ)

로그 우도의 핵심은 $\log L(\mu)=\sum_i\sum_j\sum_k x_{ijk}\log\mu_{ijk}-\sum_i\sum_j\sum_k\mu_{ijk}-\sum_i\sum_j\sum_k\log(x_{ijk}!)$이고, 식(1)에 (2)를 대입하면, 로그 확률 값을 아래와 같이 정의할 수 있다.

$$\begin{aligned}\log f(x|\mu)=&\lambda x_{+++}+\sum_i\lambda_i x_{i++}+\sum_j\lambda_j x_{+j+}+\sum_k\lambda_k x_{++k} \quad \cdots(4)\\&+\sum_i\sum_j\lambda_{ij+}x_{ij+}+\sum_i\sum_k\lambda_{i+k}x_{i+k}\\&+\sum_j\sum_k\lambda_{+jk}x_{+jk}+\sum_i\sum_j\sum_k\lambda_{ijk}x_{ijk}\\&-\sum_i\sum_j\sum_k\exp(\lambda+\cdots+\lambda_{ijk})\\&-\sum_i\sum_j\sum_k\log(x_{ijk}!).\end{aligned}$$

여기에서 $x_{+++} = \sum_i \sum_j \sum_k x_{ijk}$ 이고, x_{i++}, x_{+j+}, x_{++k}, x_{ij+}, x_{i+k}, x_{+jk} 은 분할표에서 각각 다른 부분합을 의미하고, 이것들 또한 비슷하게 정의될 수 있다. 포아송 분포는 지수족에 속하기 때문에, 모수의 추정량은 충분통계량이다. 이 포화모형에서 $\{x_{ijk}\}$는 $\{\lambda_{ijk}^{XYZ}\}$의 계수이고, 여기에서 데이터의 축소는 있을 수 없다. 좀 더 간단한 모형을 보기 위해, 어떤 모수들은 0이라 하고 식(4)를 정리하자. 식(4)에서 0으로 축소되는 몇 모형들에 대하여, 데이터는 상응하는 충분통계량에 근거하여 축소된다. 예를 들어, $\lambda_{ij} = \lambda_{jk} = \lambda_{ik} = \lambda_{ijk} = 0, \ \forall i,j,k$ 인 상호 독립 모형에 대하여 $(\lambda_i, \lambda_j, \lambda_k)$의 충분통계량은 $(x_{i++}, x_{+j+}, x_{++k})$이다.

보통, 주변 확률분포가 주어지면, X의 조건부 분포는 아래와 같이 표현될 수 있다.

$$P(x|S) = \frac{C}{\prod_i \prod_j \prod_k x_{ijk}!}. \quad \cdots(5)$$

여기에서 S는 충분통계량의 집합을 나타내고 C는 분포를 적절하게 만들어주는 상수를 나타낸다. 예를 들어, 상호독립모형에 대해, $C = \prod_i x_{++}! \prod_j x_{+j+}! \prod_k x_{++k}! / x_{+++}!$이다. Baglivo et al.(1988)에서는 다른 주변제약이 있는 다른 모형들에 대한 C 표현을 담고 있다.

제 3 장 분할표 분석을 위한 기존 방법론

3.1 능형(ridge) 벌점화 우도 방법

θ에 대한 추정량 $\hat{\theta}$의 평균 제곱 오차(mean squared error, MSE)는 $MSE(\hat{\theta}) = E(\hat{\theta}-\theta)^2 = Var(\hat{\theta}) + [E(\hat{\theta})-\theta]^2$ 로 정의된다. 즉, 평균제곱오차는 추정량의 분산과 편의의 제곱의 합으로 이루어지는데, 가우스 마코프 정리 (Gauss Markov theorem)에 의하면 회귀 문제에서 최소제곱추정량은 선형 불편 추정량 중에서 분산의 최솟값을 가진다. 불편성에서 벗어나서 생각해보면 편의가 있지만 분산은 작아서 전체 평균제곱오차 측면에서 최소제곱추정량보다 더 좋은 추정량이 존재할 수 있다. 이것에 기반한 것이 축소 추정(shrinkage estimation)이다. 능형 회귀와 라소회귀는 최소제곱추정에 대한 축소 추정법이라고 할 수 있다.

Hoerl와 Kennard(1970)가 제안한 능형회귀는 설명변수들이 비직교적일 때 사용될 수 있는 추정방법으로 공선성의 탐색과 회귀계수의 추정을 동시에 처리해 주는 방법이다. 그리고 추정에 사용된 자료의 작은 변화에 추정량이 크게 변하지 않는다는 장점을 가진다(Hoerl and Kennard, 1970). 공선성이 높을 경우 자료의 교란 상태가 추정된 회귀계수에 큰 영향을 미치는데 능형회귀의 경우 통상적 최소제곱추정 값보다 더 로버스트(robust)한 추정 값을 제공한다. 또한 이 방법은 자료에서의 작은 변화에 대한 OLS 추정 값의 안정성과 민감성의 정도를 나타내는 데에도 사용된다.

$$\sum_i (y_i - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \quad \cdots(6)$$

능형회귀 추정량은 제약조건 $\sum_{j=1}^{p} \beta_j^2 \le t^2$ 하에서 아래 식과 같다.

$$\hat{\beta}^{ridge} = \arg\ \min(\beta) \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \qquad \cdots(7)$$

라그랑즈 상수법(Lagrange multiplier)에 의하면 식(7)은 아래와 동치가 된다.

$$\hat{\beta}^{ridge} = \arg\ \min(\beta) \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \qquad \cdots(8)$$

여기에서 $t \geq 0$ 이고 $\lambda \geq 0$ 이다.

능형회귀는 $p > n$ 일 때 최소제곱추정량을 계산하기 위해 고안되었다. x_{ij}를 $x_{ij} - \overline{x_i}$로, y_{ij}를 $y_{ij} - \overline{y_i}$로 대체하면 상수항이 없는 회귀모형을 고려할 수 있다. X를 $n \times p$ 행렬이라고 하면 통상적인 최소제곱추정량은 $\hat{\beta}^{ls} = (X^T X)^{-1} X^T y$ 로 주어진다. $p < n$ 인 경우에는 $(X^T X)^{-1}$ 가 존재하지 않는다. 따라서 정규 방정식 $(X^T X)\hat{\beta} = X^T y$ 의 해는 유일하지 않다는 것을 알 수 있다. 능형회귀 추정량은 앞의 정규방정식에서 역행렬이 존재하도록 $X^T X$ 에 대각행렬 λI를 더한 것이다. 능형 회귀의 추정량은 아래와 같다.

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y. \quad \cdots(9)$$

설명변수들이 서로 직교하는 경우, 즉 $X^T X = I$ 일 때는 아래와 같이 표현된다.

$$\hat{\beta}^{ridge} = \frac{\hat{\beta}^{ols}}{1 + \lambda}. \quad \cdots(10)$$

$\lambda \geq 0$ 이므로 능형회귀 추정량 $\hat{\beta}^{ridge}$는 최소제곱추정량 $\hat{\beta}^{ols}$의 축소된 추정량

이라고 할 수 있다. $\lambda=0$ 일 때 능형회귀 추정량은 최소제곱 회귀추정량과 같은 값이 된다. 능형회귀 추정량은 λ 값에 따라 달라지므로 λ 값의 선택이 중요한 문제이며 λ 값을 선택하는 기준은 다양하다. λ 값은 추정량의 효율성을 높일 수 있도록 추정량의 평균제곱오차가 작게 되는 것을 선택해야 한다. 이는 분산 팽창 인자(variance inflation factor, VIF) 혹은 상태지수(condition index)를 같이 검토하여 결정하게 된다. 본 연구에서는 모수 λ를 최소 제곱근 하나남기기 교차 검증(root leave-one-out cross validation, RLOOCV) 오분류율을 최소화함으로써 선택한다. RLOOCV은 거의 모든 데이터로 모형화를 하기 때문에 편차가 작아 모형의 안정성을 위해 많이 선택된다. 능형회귀 추정량은 편차는 있지만 분산을 줄이면서 전반적인 예측오차를 줄이는 방법이 된다.

3.2 라소(lasso) 벌점화 우도 방법

능형회귀는 축소 추정량을 제공하지만 변수 선택은 하지 않는다. 그렇기 때문에 고차원 자료의 경우 최종 모형에 대한 해석이 어려운 문제점이 있다. Tibshirani(1996)이 제안한 라소회귀는 축소와 변수선택을 통해 예측력을 향상시키는 동시에 최종 모형에 대한 해석을 쉽게 하는 방법이다. 라소 추정량은 제약조건 $\sum_{j=1}^{p}|\beta_j| \le t$ 하에서 아래의 식과 같다.

$$\widehat{\beta^{lasso}} = \arg\ \min(\beta)\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2. \qquad \cdots(11)$$

능형회귀와 마찬가지로 라그랑즈 상수법에 의하면 식(11)은 아래의 식과 동치가 된다.

$$\widehat{\beta^{lasso}} = \arg\ \min(\beta)\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}. \qquad \cdots(12)$$

능형회귀와 라소회귀의 차이는 벌점이 l_2-노음(norm) $\sum_{j=1}^{p}\beta_j^2$ 에서 l_1-노음(norm)인 $\sum_{j=1}^{p}|\beta_j|$로 바뀌었다는 것이다. 라소는 설명력이 없는 설명변수들의 계수를 0으로 추정함으로써 자동적인 변수선택이 이루어진다. 또한 최적 부분 집합 선택법은 특정 입력 변수를 선택하거나 제거하는 불연속적인 변수 선택인 반면, 라소는 축소를 통해 연속적으로 변수를 선택하는 장점을 가진다.

3.3 엘라스틱넷(elastic net) 벌점화 우도 방법

엘라스틱넷의 벌점함수는 라소와 능형회귀의 l_1-노음과 l_2-노음 벌점을 절충안으로 사용하고 있다(Zou and Hastie, 2005). 이 같은 조합은 능형회귀의 정규화 특성을 지니면서, 라소와 같이 일부 계수 추정 값을 0까지 축소시켜 변수선택을 제공한다. 즉, 엘라스틱넷은 능형 회귀와 라소 회귀의 절충으로서, 상관관계가 있는 변수들 중에서 하나의 변수만을 선택하는 라소의 단점을 보완하기 위해 제안되었다. 엘라스틱넷은 l_1-노음과 l_2-노음을 결합함으로써 상관관계가 있는 변수들은 모두 선택하게 하고, 이것을 그룹화 효과(grouping effect)라고 부른다. 또한 높은 상관도를 가진 추정 변수가 데이터의 크기에 비해 많을 때 유용하다. 엘라스틱넷의 추정량은 아래와 같다.

$$\widehat{\beta^{elasticnet}} = \arg\ \min(\beta)\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_2\sum_{j=1}^{p}\beta_j^2 + \lambda_1\sum_{j=1}^{p}|\beta_j|\right\}. \qquad \cdots(13)$$

설명변수가 두 개인 회귀분석의 경우 α와 λ 값들이 주어졌을 때 두 변수의

추정계수를 각각 $\widehat{\beta_1}(\alpha,\lambda)$와 $\widehat{\beta_2}(\alpha,\lambda)$라 하자. 그러면 어떤 상수 $M>0$에 대하여 다음 부등식이 성립한다.

$$|\widehat{\beta_1}(\alpha,\lambda)-\widehat{\beta_2}(\alpha,\lambda)<\frac{\sqrt{n}M}{\alpha\lambda}\sqrt{2(1-r_{12})}|. \quad \cdots(14)$$

여기에서 r_{12}는 두 변수간의 상관계수를 나타낸다. 따라서 두 변수의 상관계수가 1에 가까울수록 두 계수의 차이는 0이 된다. 즉, 두 변수간의 상관계수가 크면 대응되는 추정량들의 값은 거의 동일하게 된다.

3.4 베이지안 부분집합 회귀(BSR) 방법

Liang et al.(2013)은 고차원에서 small-n-large-p의 경우에 전통적인 방법의 문제점을 극복하고자 최소 EBIC와 동일한 최대 사후모형인 새로운 베이지안 부분집합 회귀(BSR) 모형을 제안했다. 이 장에서는 SAMC를 이용한 BSR 방법에 대해 소개한다.

포아송 로그선형모형에서 D^n의 부분집합 모형이라 하자. 이때 $|\xi_n|$를 ξ_n 모형의 크기, ξ_n의 모수를 $\beta_{\xi_n}=\{\beta_{1,}\cdots,\beta_{|\xi_n|}\},\beta_m=\{(\lambda,\lambda_{(i)},\lambda_{(j)},\lambda_{(k)},\lambda_{(ij)},\lambda_{(ik)},\lambda_{(jk)},\lambda_{(ijk)}),$ $i=1,\cdots,I;\ j=1,\cdots,J;\ k=1,\cdots,K\},\ m=1,\cdots,|\xi_n|$ 이라 하자.

실제 모형이 0인 값들을 많이 포함하는 자료이고 $\lim_{n\to\infty}\sum_{i=1}^{P_n}|\beta_i|<\infty$라는 조건을 만족한다고 가정하고, 어떤 양의 상수 C_0에 대해서 $\sigma_{\xi_n}^2=\frac{1}{2\pi}e^{\frac{C_0}{|\xi_0|}}$라고 설정한다면, β_{ξ_n}의 사전분포는 $|\log(\pi(\beta_{\xi_n}))|=|-\frac{C_0}{2}-\pi e^{-\frac{C_0}{|\xi_n|}\sum_{i=1}^{|\xi_n|}\beta_i^{*2}}|\le\frac{C_0}{2}+\pi\sum_{i=1}^{P_n}\beta_i^2<\infty$가 된다. 어떤 모형 ξ_n에 대해서든 β_{ξ_n}의 사전정보는 충분히 큰 n에 대해서는 무시될 수 있다는 것을 암시한다. 더 큰 C_0은 β_{ξ_n}의 고른 사전분포를 유도하기 때문에

C_0을 선택할 때 보통 더 큰 값이 선호된다. 그러나 이 연구에서는 Liang et al.(2013)이 제안한 정확한 사후분포 대신 대략적인 사후분포로부터 표본추출하는 것을 선택했고, C_o값을 선택하는 문제가 사라졌다.

포아송 로그선형모형에서 변수선택을 위해 Liang et al.(2013)이 제안한 ξ_n의 로그 사후분포를 이용한다. ξ_n의 사전분포로 $\pi(\xi_n)=\nu_n^{|\xi_n|}(1-\nu_n)^{P_n-|\xi_n|}$이라 정의한다. 즉, 각각의 변수들은 부분집합 모형을 위해표본추출 될 다른 변수들로부터 독립적인 사전 성공확률 ν_n을 갖는다. 우도함수의 형태를 BIC형태로 유도해 나타내면 아래와 같다.

$$\log f(y_n|\beta_{\xi_n},\xi_n,X_n)\approx logf(y_n|\widehat{\beta_{\xi_n}},X_n)-\frac{n}{2}(\beta_{\xi_n}-\widehat{\beta_{\xi_n}})'J_n(\widehat{\beta_{\xi_n}})(\beta_{\xi_n}-\widehat{\beta_{\xi_n}}). \quad \cdots(15)$$

여기에서 $y_n=(y_1,y_2,\cdots,y_n)$이고 $X_n=(x_1,x_2,\cdots,x_n)$, $\widehat{\beta_{\xi_n}}$은 β_{ξ_n}의 최대우도추정량(MLE), $J_n=-\frac{1}{n}\frac{\partial^2\log f(y_n|\beta_{\xi_n},\xi_n,X_n)}{\partial\beta_{\xi_n},\partial\beta_{\xi_n}'}|_{\beta_{\xi n}=\widehat{\beta_{\xi_n}}}$은 β_{ξ_n}추정량들의 공분산의 역행렬이다.

사전분포에 대해서는 $\Pi(\beta_{\xi_n})\approx\Pi(\widehat{\beta_{\xi_n}})+(\beta_{\xi_n}-\widehat{\beta_{\xi_n}})'\frac{\partial\Pi(\beta_{\xi_n})}{\partial\beta_{\xi_n}}|_{\beta_{\xi_n}=\widehat{\beta_{\xi_n}}}$ 이다. 따라서

$$\begin{aligned}
f(y_n|\xi_n,X_n)&=\int f(y_n|\beta_{\xi_n},\xi_n,X_n)\pi(\beta_{\xi_n})d\beta_{\xi_n}\\
&\approx\int\exp\left\{\log f(y_n|\widehat{\beta_{\xi_n}},\xi_n,x_n)-\frac{n}{2}(\beta_{\xi_n}-\widehat{\beta_{\xi_n}})^TJ_n(\widehat{\beta_{\xi_n}})(\beta_{\xi_n}-\widehat{\beta_{\xi_n}})\right\}\\
&\quad\times\left\{\pi(\widehat{\beta_{\xi_n}})+(\beta_{\xi_n}-\widehat{\beta_{\xi_n}})'\frac{\partial\pi(\beta_{\xi_n})}{\partial\beta_{\xi_n}}|_{\beta_{\xi_n}=\widehat{\beta_{\xi_n}}}\right\}d\beta_{\xi_n}\\
&\approx f(y_n|\widehat{\beta_{\xi_n}},\xi_n,X_n)\pi(\widehat{\beta_{\xi_n}})\times\int\exp\left\{-\frac{n}{2}(\beta_{\xi_n}-\widehat{\beta_{\xi_n}})^TJ_n(\widehat{\beta_{\xi_n}})(\beta_{\xi_n}-\widehat{\beta_{\xi_n}})\right\}d\beta_{\xi_n}\quad\cdots(16)\\
&=f(y_n|\widehat{\beta_{\xi_n}},\xi_n,X_n)\pi(\widehat{\beta_{\xi_n}})\frac{(2\pi)^{\frac{|\xi_n|}{2}}}{n^{\frac{|\xi_n|}{2}}J_n(\widehat{\beta_{\xi_n}})^{\frac{1}{2}}}.
\end{aligned}$$

이것은

$$\log\pi(\xi_n|D^n) = C + \log\{\pi(\xi_n)f(y_n|\xi_n, X_n)\} \quad \cdots(17)$$
$$\approx C + \log\pi(\xi_n) + \log f(y_n|\widehat{\beta_{\xi_n}}, \xi_n, X_n) + \log\pi(\widehat{\beta_{\xi_n}}) - \frac{|\xi_n|}{2}\log(n)$$
$$+ \frac{|\xi_n|}{2}\log(2\pi) - \frac{1}{2}\log|J_n(\widehat{\beta_{\xi_n}})|.$$

를 의미한다. 여기에서 C는 상수, $J_n(\widehat{\beta_{\xi_n}})$은 추정량들의 정보행렬이다. 그리고 더 나아가 사전확률 ν_n이 어떤 모수 γ에 대해 $\nu_n = \frac{1}{1+P_n^{\gamma}\sqrt{2\pi}}$ 값을 갖는다면 근사된 로그 사후분포를 갖는다.

$$\log\pi(\xi_n|D^n) \approx C + \log f(y_n|\widehat{\beta_{\xi_n}}, \xi_n, X_n) - \frac{|\xi_n|}{2}\log(n) - |\xi_n|\gamma log(P_n). \quad \cdots(18)$$

여기에서 $\widehat{\beta_{\xi_n}}$은 β_{ξ_n}의 MLE이고, $\log f(y_n|\widehat{\beta_{\xi_n}}, \xi_n, X_n)$는 y_n의 확률값으로 $f(y_n)$를 어떻게 표현하느냐에 따라 형태가 달라지는 일반화 선형모형이다. 또한, $-\frac{|\xi_n|}{2}\log(n) - |\xi_n|\gamma log(P_n)$은 임의의 모수 γ와 함께 설정하고 들어가는 항이다. 식(18)은 최종 사후분포 식으로, 이 사후확률을 최대화 하는 것은 EBIC(= $-2\log f(y_n|\widehat{\beta_{\xi_n}}, \xi_n, X_n) + |\xi_n|\log(n) + 2|\xi_n|\gamma log(P_n)$)를 최소화하는 것과 같다.

MLE $\widehat{\beta_{\xi n}}$의 계산을 쉽게 하기 위해, 모형 크기인 $|\xi_n|$에 상한 경계를 둘 필요가 있다. 예를 들어, $|\xi_n| > K_n$이라고 하면 여기서 K_n은 표본 크기 n에 의존한다. 만약 $|\xi_n| > n$이라면 행렬 $\widehat{X_{|\xi_n|}}'\widehat{X_{|\xi_n|}}$이 비정칙행렬이 되는데, 여기에서 $\widehat{X_{|\xi_n|}} = (x_1^*, x_2^*, \cdots, x_{|\xi_n|}^*)$는 모형 ξ_n의 설계행렬이다. 이 $|\xi_n| > n$ 경계에서, 모형 ξ_n의 사전분포는 $\pi(\xi_n) \propto \nu_n^{|\xi_n|}(1-\nu_n)^{P_n - |\xi_n|} I[|\xi_n| < K_n]$이고, 로그-사후분포는 아래와 같이

정의된다.

$$\log\pi(\xi_n|D^n) \approx \begin{cases} C+\log f(y_n|\widehat{\beta}_{\xi_n},\xi_n,X_n)-\dfrac{|\xi_n|}{2}\log(n)-|\xi_n|\gamma log(P_n) & (\text{if } |\xi_n|<K_n) \\ -\infty \ \ (otherwise) & \end{cases} \cdots(19)$$

.

위의 로그-사후분포를 음으로 표현하면 EBIC로 거의 비슷하게 축소하기 때문에, 이것은 BSR이라고 불린다.

본 연구에서는 Wilks 우도비 통계량을 이용하여 ξ_n의 로그 사후분포로 다음과 같은 BSR 모형을 정의한다.

$$\log\pi(\xi_n|D^n) \approx \sum_{\xi_n}[x_{ijk}\log(x_{ijk}/\widehat{\mu_{ijk}})]-\frac{|\xi_n|}{2}\log(n)-|\xi_n|\gamma log(P_n). \ \cdots(20)$$

여기서 $\widehat{\mu_{ijk}}$은 모형 ξ_n하에서 μ_{ijk}의 MLE이다. 식(20)에서의 사전 초기하모수 γ의 값은 다음과 같은 과정에 의해 결정될 수 있다.

(a) 일련의 서로 다른 γ를 정한다.
(b) 각각의 γ에 대하여, BSR 방법을 짧게 반복 시행한다.
(c) BSR 사후분포의 MAP(maximum a posteriori) 모형의 크기와 일치하는 $\Pi(|\xi_n||D^n,\gamma)$의 최빈값을 가지는 γ의 최솟값을 선택한다.

본 연구는 BSR 사후분포 (20)으로부터 표본 β_{ξ_n}을 생성시키기 위해 SAMC 알고리즘을 이용한다.

$$f(x)=\frac{1}{Z}\psi(x), \quad x\in\chi. \ \cdots(21)$$

위의 분포로부터 표본을 뽑고 싶다고 가정하자. 여기에서 Z는 정규화 상수이고 χ는 표본공간이다. BSR에 SAMC를 적용하기 위해 $U(x)=-\log(\psi(x))$로 두고, 이것을 식(21) 분포의 에너지함수라고 한다. 표본공간 χ가 에너지함수에 따라 $E_1=\{x:U(x)\le u_1\}$, $E_2=\{x:u_1<U(x)\le u_2\}$, $\cdots$, $E_{m-1}=\{x:u_{m-2}<U(x)\le u_{m-1}\}$, $E_m=\{x:U(x)>u_{m-1}\}$으로 표기되는 m개의 소구역으로 나뉜다고 가정한다. 여기에서 $u_1,\cdots,u_{m-1}$은 미리 지정된 상수들이다. $i=1,\cdots,m$에 대하여 $w_i=\int_{E_i}\psi(x)dx$라고 두고 추정한다. 여기에서 $w_i>0$인데, 모든 영역에 확률 값이 존재하기 때문에 확률 값이 0인 영역이 없다는 것을 보인다. SAMC는 실험분포 $f_w(x)\propto\sum_{i=1}^{m}\frac{\pi_i\psi(x)}{w_i}I(x\in E_i)$로부터 표본추출을 하려고 한다. 여기서 $\sum_{i=1}^{m}\frac{\pi_i\psi(x)}{w_i}$은 부분 영역별로 가중치를 두어 혼합분포를 만든 것이다. π_i는 모든 i들에 대하여 $\pi_i>0$이고 $\sum_{i=1}^{m}\pi_i=1$인 미리 지정된 상수들이다.

Liang et al.(2007)에서 $\pi=(\pi_1,\cdots,\pi_m)$은 소구역에 대해 분석자가 원하는 확률분포라고 나타낸다. 만약 $w_1,\cdots,w_m$이 잘 추정될 수 있다면, $f_w(x)$로부터 표본추출하는 것은 랜덤 워크(random walk)가 된다. 이것은 소구역 공간에서 각 소구역을 하나의 점으로 보고, 각 소구역은 π_i에 비례하는 빈도로 표본추출 된다는 것을 말한다. π_i를 직접 수정하는 것이 아니라 w_i를 수정하여 각 영역별로의 빈도수를 조절할 수 있는데, 상대 빈도수가 π_i와 비례한다. 이런 이유로 w_i가 잘 추정되어 표본공간이 적절하게 분할된다면 국소 트랩(local trap)문제가 근본적으로 해결된다. 타겟 분포를 조정하는 적응 메트로폴리스 (adaptive Metropolis, Haario, Saksman, and Tamminen, 2001)와 달리 SAMC는 변함없는 타겟 분포를 조절한다. 이런 의미에서 SAMC는 동적 중요표본추출(dynamic importance sampling) 알고리즘처럼 보이기도 한다.

$\widehat{\theta_{ti}}$를 t반복으로부터 얻어진 $\log(\frac{w_i}{\pi_i})$의 추정량을 나타내도록 하자. 여기에서 $\theta_t=(\theta_{t1},\cdots,\theta_{tm})$이다. 어떤 $\zeta\in(1,2)$에 대하여

$$\text{(i)}\ \sum_{i=1}^{\infty} a_t = \infty \qquad\qquad \text{(ii)}\ \sum_{t=1}^{\infty} a_t^{\zeta} < \infty \qquad \cdots(22)$$

위의 조건들을 만족하면서 증가하지 않는 양의 상수를 $\{a_t\}$로 두자. 본 연구에서는 $a_t = \frac{t_0}{\max(t_0, t)}$, $t = 1,2,\cdots$를 사용한다. 여기에서 t_0는 미리 결정된 임의의 양수로, 그 선택은 표본공간의 복잡한 정도에 의존할 수 있다. 보통 m의 큰 값은 t_0의 큰 값과 연관된다.

$w = (w_1, \cdots, w_m)$의 크기 변화에 대해서 $f_w(x)$가 변함이 없기 때문에 θ_t의 범위는 유한한 범위 Θ로 제한될 수 있다. 하지만 현실적인 문제로서 그 범위를 $\Theta \in \mathbb{R}^m$으로 설정하는 것과 같다. 이 문제를 해결하기 위해 θ가 w와만 관련 있게 만들기 위해 π가 전 소구역에서 균일하게 한다.

SAMC 알고리즘:

(a) (표본추출) 한번의 Metropolis-Hastings 업데이트에 의해 표본 x_t를 생성한다. 여기서 x_t의 타겟 분포는 아래와 같다.

$$f_{\theta_t}(x) \propto \sum_{i=1}^{m} \frac{\psi(x)}{e^{\theta}_{t_i}} I(x \in E_i). \quad \cdots(23)$$

(b) (가중치 업데이트) $\theta^* = \theta_t + a_{t+1}(\widehat{e_t} - \pi)$를 설정한다. 여기에서 $\widehat{e_t} = (\widehat{e_{t,i}}, \cdots, \widehat{e_{t,m}})$이고 만약 $x_t \in E_i$라면 $\widehat{e_{t,i}} = 1$이고 그렇지 않으면 0이다. 그리고 만약 $\theta^* \in \Theta$이면 $\theta_{t+1} = \theta^*$라고 설정하고 그렇지 않으면 $\theta_{t+1} = \theta^* + C^*$로 설정한다. 여기에서 $\boldsymbol{C^*} = (C^*, \cdots, C^*)$는 조건 $\theta^* + C^* \in \Theta$를 만족하는 임의의 벡터이다.

Liang et al(2007)에서는 $t \to \infty$일 때

$$\theta_{ti} \to \begin{cases} Const + \log(w_i) - \log(\pi_i + \overline{\pi_0}), & \text{if } E_i \neq \varnothing \\ -\infty, & \text{if } E_i = \varnothing \end{cases}. \quad \cdots(24)$$

이라는 것을 보였다. 여기에서 $\overline{\pi_0}=\sum_{j\in\{i:E_i=\varnothing\}}\frac{\pi_j}{m-m_0}$ 이고, m_0는 비어있는 소구역의 개수, $Const$는 임의의 상수를 나타낸다. $w_i=0$일 때, 소구역 E_i는 비어있는 소구역이라고 불린다.

$(x_1,w_1),\cdots,(x_N,w_N)$이 SAMC의 (a)표본추출 단계에서 얻어진 표본들의 집합을 나타낸다고 하자. 여기에서 $w_t=\sum_{i=1}^{m}e^{\theta_{ti}}I(x_t\in E_i)$이다. $y_1,\cdots,y_{N'}$은 $x_1,\cdots,x_N$중에서 구분된 표본들을 나타낸다고 하자.

$$P(Y=y)=\frac{\sum_{t=1}^{N}w_tI(x_t=y)}{\sum_{t=1}^{N}w_t},\ y\in\{y_1,\cdots,y_{N'}\}. \quad \cdots(25)$$

만약 위의 가중치 확률 값으로 랜덤표본 Y를 만들어낸다면 Y는 밀도함수 $f(x)$에 대하여 점근적으로 분포된다. 이 성질은 SAMC가 중요도 표본추출 알고리즘으로 사용될 수 있다는 것을 암시한다. 어떤 적분 가능한 함수 $h(x)$에 대하여, 기댓값 $E_fh(x)=\int h(x)f(x)$는 아래 식에 의해 추정될 수 있다.

$$\widehat{E_fh(x)}=\frac{\sum_{t=1}^{N}w_th(x_t)}{\sum_{t=1}^{N}w_t}. \quad \cdots(26)$$

중요도 표본추출 추정량이 수렴하는 것과 같은 이유로 $n\to\infty$일 때 $\widehat{E_fh(x)}\to E_fh(x)$이 성립한다.

메트로폴리스-헤스팅스(MH, Metropolis et al., 1953) 알고리즘 같은 전통적인 MCMC 알고리즘과 비교하여 SAMC는 표본공간 탐색에 있어서 중요한 이점을 갖는다. 이것은 SAMC의 자기조절(self-adjusting) 구조 때문이다. 만약 소구역이 t 반복 째에 방문되면, θ_t는 그에 맞춰 업데이트 될 것인데 이 소구역

은 그 다음 반복에서 재방문 될 확률이 더 작아진다. 이것이 SAMC 알고리즘에 있어서 π 값의 역할이기도 하다. 자기조절 구조는 SAMC가 국소 트랩의 영향을 받지 않고, 특히 고차원 공간의 표본추출에 적합하게 한다.

일반화선형모형(GLM)의 사후분포로부터 모의실험에 SAMC를 적용하기 위해 MH 과정인 생성(birth), 소멸(death), 그리고 교환(exchange)을 명시한다. 모의실험의 수렴을 가속화하기 위해, 각각의 예측변수에 똑같은 가중치를 주지 않고 구역별로 뽑힐 확률을 다르게 한다. 각각의 예측 변수의 가중치를 결정하기 위해, 절편항 하나와 예측변수 하나로 구성된 P_n개의 모형들의 편차를 평가한다. 가중치의 중요도를 아래와 같이 표현할 수 있다.

$$\rho_i = \max\left\{0.0001, \exp\left\{\frac{-(d_i - d_{\min})}{(S\sigma_d) - d_0}\right\}\right\}, \quad i = 1, \cdots, P_n. \quad \cdots(27)$$

여기에서 $d_1, \cdots, d_{P_n}$은 P_n개의 모형들의 편차를 나타내고 σ_d는 $d_1, \cdots, d_{P_n}$의 표준편차를 나타내며, $d_0 > 0$ 그리고 $S > 0$은 사용자가 지정하는 모수들이다.

그러므로 예측변수가 더 작은 편차를 가질수록 더 큰 가중치를 갖는다. 이것은 작은 편차를 가질수록 그 변수가 뽑힐 확률을 더 크게 주겠다는 것을 의미한다. 생성과 소멸 단계에서 암시되듯이, 모수 S는 예측변수들의 상대적인 가중치 규모를 조절하고, 모수 d_0은 제안한 것을 원상태로 되돌릴 수 있게 한다. 기본 설정은 $S = 5$ 그리고 $d_0 = 0.1$이다.

ξ_t를 현재 모형이라고 할 때, ξ_n은 사후분포 $\pi(\xi_n | D^n)$에서 뽑힌 모형을 나타낸다. 이런 자료에서, 앞서 언급한 사후분포로부터 뽑힌 포괄적인 모형을 나타내기 위해 ξ을, 반복 t에서 SAMC에 의해 뽑힌 모형을 나타내기 위해 ξ_t를 쓰기로 한다. 모형 ξ_t에 포함된 예측변수들 집합을 $S_t = \{i : x_i \in \xi_t\}$로 나타내고, 모형 ξ_t로부터 제외되는 예측변수들 집합을 $S_t^c = \{i : x_i \notin \xi_t\}$로 나타낸다고 하자.

(a) 생성 이동(birth move). 예측변수, 즉 x_i는 아래의 확률로 S_t^c집합에서 선

택되고, 그런 다음 새 모형은 현재 모형에서 x_i를 포함시켜 형성된다.

$$P(x_i|birth,\xi_t) = \frac{\rho_i}{\sum_{k\in S_t^c}\rho_k}. \quad \cdots(28)$$

(b) 소멸 이동(death move). 예측변수, 즉 x_j는 아래의 확률로 S_t집합에서 선택되고, 그런 다음 새 모형은 현재 모형에서 x_j를 제거하여 형성된다.

$$P(x_j|death,\xi_t) = \frac{1-\rho_j}{\sum_{k\in S_t}(1-\rho_j)}. \quad \cdots(29)$$

(c) 교환 이동(exchange move). 예측변수, 즉 x_i는 확률 (28)인 S_t^c집합으로부터 선택되고, 또 다른 예측변수인 x_j는 확률 (29)인 S_t집합으로부터 선택된다. 그리고 그런 다음 새 모형은 현재 모형에서 x_j를 x_i로 교환하여 형성된다.

소멸과 교환 이동은 상수항만 있는 영모형(null-model)에서 수행될 수 없고, 생성 이동은 최대 크기 모형에서 수행될 수 없기 때문에, ξ_t 모형에 포함된 예측 변수들의 개수인 $|\xi_t|$ 조건에서 세 가지 이동을 위해 아래의 확률을 명시한다.

$$\begin{cases} P(birth||\xi_t|=0)=1 \\ P(birth|0<|\xi_t|<K_n) = P(death|0<|\xi_t|<K_n) = P(exchange|0<|\xi_t|<K_n) \\ P(death||\xi_t|=K_n) = P(exchange||\xi_t|=K_n) = \frac{1}{2} \end{cases}. \quad \cdots(30)$$

여기에서 $K_n < P_n$은 분석자에 의해 고려된 최대 모형 사이즈를 나타낸다.

(28), (29), (30)에 주어진 교환 이동을 위한 전이확률비는 아래처럼 쓰여 질 수 있다.

$$\frac{T(\xi^{*}\rightarrow\xi_t)}{T(\xi_t\rightarrow\xi^{*})}=\frac{P(x_j|birth,\xi^{*})P(x_i|death,\xi^{*})P(exchange\|\xi^{*}|)}{P(x_i|birth,\xi_t)P(x_j|death,\xi_t)P(exchange\|\xi_t|)}. \quad \cdots(31)$$

생성 이동과 소멸 이동의 전이확률비도 비슷하게 쓸 수 있다.

고차원 분할표에서 SAMC의 효율적인 실행을 위해 몇 가지 주의해야할 문제가 있는데, 표본공간 분할과 π의 결정, 그리고 수렴진단 부분이다. 표본공간은 로그 사후밀도함수의 음의 표현인 에너지 함수 또는 모형의 크기에 따라 분할될 수 있다. 본 연구에서는 에너지 함수에 따라 표본공간을 분할하기로 결정했고, 이는 높은 사후확률 모형들에 초점을 맞춘 표본추출을 할 수 있게 해준다. 만약 모형의 크기에 따라 표본공간을 분할하기로 결정한다면, 다른 크기 모형들의 표본추출 빈도를 조절할 수 있을 것이다. 하지만 고차원 분할표의 표본공간의 거대한 크기를 고려해볼 때, 모형의 크기에 따라 분할하는 것보다 에너지 함수에 따라 표본공간을 분할하는 것이 더 효율적으로 보인다.

에너지 함수 $U(\xi)$가 주어지면, 표본공간은 다음과 같이 분할될 수 있다.

$$\begin{aligned} &E_1=\{\xi: U(\xi)\le u_1\},\ E_2=\{\xi: u_1<U(\xi)\le u_2\},\ \cdots, \\ &E_{m-1}=\{\xi: u_{m-2}<U(\xi)\le u_{m-1}\},\ E_m=\{\xi: U(\xi)>u_{m-1}\}. \end{aligned} \quad \cdots(32)$$

여기에서 $u_i=u_1+(i-1)\triangle u,\ \ i=1,\cdots,m-1$ 이고, u_i는 미리 지정된 수이다. 본 연구에서는 $\triangle u=1$로, E_1은 비어있게 u_1을 매우 작은 숫자가 되게 설정하고, 소구역 E_m으로 분할된 모형들은 연구의 관심사가 되지 않도록 m을 매우 큰 숫자가 되게 설정한다. 이는 u_1과 m의 선택이 에너지의 차원에서 예상된 표본공간을 SAMC에 의해서 잘 탐색되게 만든다.

주변 포함확률은 고차원 분할표에서의 변수선택에 있어서 중요한 역할을 한

다. 주변 포함확률은 변수들의 중요도를 결정하기 때문이다. 어떤 변수가 실제로 중요한지 사전정보가 없으므로 각각의 예측변수를 합리적인 빈도로 뽑히게 만든다. 이를 위해, 원하는 표본추출 분포가 $\pi_1 = \pi_2 = \cdots = \pi_m = \frac{1}{m}$ 처럼 균일분포가 되게 한다. 만약 낮은 에너지 소구역으로 편향된 표본추출을 한다면, 어떤 예측변수들은 뽑힐 기회가 거의 없을 것이고 주변 포함 확률의 결과 예측 값은 편차가 커질 것이다.

SAMC의 수렴은 다중 실행에 근거하여 진단될 수 있다. 예를 들어, 각 실행에서 나온 표본추출 빈도를 원하는 표본추출 빈도와 비교할 수 있다. 만약 실행된 표본추출 빈도가 전부 원하는 표본추출과 비슷한 패턴을 보인다면 그 실행 결과는 수렴한다고 결론 낸다. 각 실행에서 원하는 표본추출 분포가 균일분포라면 실행 종료 전에 에너지 공간에서 평평한 히스토그램이 나와야 한다. Wang and Landau(2001)에 의해 제안되었던 것처럼, 만약 각각의 소구역표본추출 빈도가 평균표본추출 빈도의 80% 미만이 아니라면 히스토그램은 평평하다고 간주될 수 있다. 이것은 실행 결과가 수렴한다는 것을 의미한다.

제 4 장 다중가설 검정 기반 변수 검사과정을 이용한 베이지안 부분집합 회귀(SVS-BSR) 변수선택 방법

비록 3.4장에서 설명한 MAP(Maximum A Posteriori)모형이 변수선택 $\hat{\xi}_{\hat{q}}$에 대한 좋은 점근선의 특성들을 제공하지만, 주어진 데이터셋에 대하여 $\hat{q}$를 어떻게 선택할 것인지는 아직 분명하지가 않다. 이 장에서 우리는 $\hat{q}$값을 결정하기 위해 다중가설 검정기반 변수 검사과정(multiple-hypothesis test-based sure variable screening, SVS)을 이용한다.

중요한 변수들이 중요하지 않은 변수들보다 더 큰 주변포함확률을 갖는 경향이 있다. $q_{1,}q_{2,}\cdots,q_{p_n}$을 P_n개의 중요한 변수들의 주변포함확률을 나타내도록 두자. 그리고 $Z_i=\Phi^{-1}(q_i),\quad i=1,2,...,P_n$ 이 상응하는 주변포함점수(marginal inclusion score, MIS)를 나타내도록 두고, 여기에서 $\Phi(\bullet)$는 표준정규분포의 누적분포함수를 나타낸다. 큰 MIS를 갖는 변수들을 찾아내기 위해, 우리는 2성분형 혼합 지수멱분포로 MIS를 모형화한다.

2성분형 혼합 지수멱분포를 위해, 밀도함수는

$g(z|\theta)=\sum_{i=1}^{2}\varpi_i\psi(z|\nu_i,\sigma_i,\alpha_i)$로 주어지는데, $\theta=(\varpi_1,\nu_1,\sigma_1,\varpi_2,\nu_2,\sigma_2)$은 분포의 모든 변수들을 포함하고, ϖ_i는 $0<\varpi_i<1$, $\sum_{i=1}^{2}\varpi_i=1$인 i번째 성분의 가중치이며, $\psi(z|\nu_i,\sigma_i,\alpha_i)=\frac{\alpha_i}{2\sigma_i\Gamma(1/\alpha_i)}\exp\left\{-(|z-\nu_i|/\sigma_i)^{\alpha_i}\right\}$이다. 여기에서 모수 $\nu_i,\sigma_i,$ 그리고 α_i는 각각 분포의 평균, 분산, 그리고 감소율을 나타낸다.

3.4장에서 설명한대로, 사후분포 $\pi(\xi|D^n)$의 표본공간은 m개의 소구역에서 에너지 함수 $U(\xi)$에 따라 분할된다고 하자. $(\xi_1,w_1),\cdots,(\xi_N,w_N)$이 SAMC의 실행에

서 나온 모형들의 집합을 나타낸다고 하고, 여기에서 $w_t = \sum_{i=1}^{m} e^{\theta_{ti}} I(x_t \in E_i)$이다. 그러면 주변 포함확률은 다음과 같은 식에 의해 추정될 수 있다.

$$\hat{q_j} = \frac{\sum_{t=1}^{N} w_t I(x_j \in \xi_t)}{\sum_{t=1}^{N} w_t}, \quad j = 1, \cdots, P_n. \quad \cdots(33)$$

여기에서 $I(\bullet)$는 지시함수이다. $N \to \infty$일 때 $\hat{q_j} \to q_j$인데, 여기에서 q_j는 변수 x_j의 실제 주변 포함확률을 나타낸다.

$\hat{q} = \Phi(z_r)$ 설정에 상응하는 절단값 z_r은 사전에 명시된 검정수준에서 중요한 변수들의 거짓 발견율(False Discovery Rate, FDR)을 조절함으로써 선택될 수 있다.

다중 가설검정은 여러 개의 가설을 동시에 검정하는 방법이다. 다중 가설 검정에서는 각각의 검정이 제1종 오류와 제2종 오류를 갖기 때문에 검정 전체의 오류율을 측정하는 것이 명확하지 않다. 검정에서는 제1종 오류와 제2종 오류의 기준으로 보통 p-value를 사용하는데, 다중 가설 검정에서 수많은 독립적인 검정을 시행하면 유의수준의 해석에 문제가 생기기 때문이다. 독립적인 검정이 많아지면 전체 검정의 유의수준이 매우 작아져 가설을 거의 기각하지 않게 되고 보수적인 검정이 된다. 이 문제를 해결하기 위해 Benjamini와 Hochberg(1995)는 다중 가설검정의 오류 단위로 FDR을 도입하였다. 변수 선택에 있어서 FDR은 유의하게 선택된 변수들 중 실제로는 유의하지 않은데 유의하다고 선택된 변수들의 기대비율로 나타내어진다.

$$FDR = E\left[\frac{\#false\ significant\ features}{\#significant\ features}\right]. \quad \cdots(34)$$

$\Lambda_r = \{z_i \geq z_r\}$ 에 대하여, FDR은 $FDR(\Lambda_r) = \frac{P_n \widehat{\varpi_1}[1 - F(z_r|\widehat{\nu_1}, \widehat{\sigma_1}, \widehat{\alpha_1})]}{\#\{z_i : z_i \geq z_r\}}$ 에 의해 추정될 수 있고, 여기에서 $\#\{z_i : z_i \geq z_r\}$은 z_r보다 더 큰 MIS를 가진 중요한 변수들의 개수를 나타내고, $F(\bullet)$은 지수멱분포의 CDF를 나타낸다.

다중 가설검정에서 FDR은 여러 개의 값이 나오고, 그 중 최소인 값을 FDR에서의 수정된 p-value, 즉 q-value라고 한다. q-value 값은 $q_r^s(z) = \inf_{\{\Lambda_r : z \in \Lambda_r\}} FDR(\Lambda_r)$ 로 정의된다. FDR이 최소라는 것은 실제로는 유의하지 않지만 유의하다고 선택된 변수들의 개수가 작다는 것을 의미한다. 따라서 우리가 용인할 수 있는 기준 값보다 작은 거짓 발견율을 선택하는 방법을 채택한다. 예를 들어, 실험수준을 0.01로 설정한다면 모든 $z \geq z_r$에 대해서 $q_r^s(z) \leq 0.01$인 z_r을 선택하는 것이다.

제 5 장 모의실험

3장과 4장에서 설명한 알고리즘의 성능을 확인하기 위해, 각각 6개의 주효과와 11개의 상호작용, 총 17개의 포아송 변수들을 포함하는 10개의 분할표를 만든 뒤, 강력하게 상호작용하는 변수들이 몇 개 있는 모의실험을 진행했다.

이 예시는 10개의 시뮬레이션 된 데이터셋을 포함한다. 데이터는 첫 번째 변수의 첫 번째 범주와 세 번째 변수의 세 번째 범주의 상호작용, 그리고 두 번째 변수의 두 번째 범주와 세 번째 변수의 세 번째 범주의 상호작용이 유의하도록 생성되었다. 각각의 데이터셋에 대하여, SAMC는 $K_n = 6$ 그리고 $\gamma = 0.85$로 실행되었다. SAMC의 각 실행은 5.1×10^5 반복했고, 여기서 첫 10,000번의 반복들은 버닝 과정을 위해 버려졌고 남은 반복들은 추론에 사용됐다. SAMC 실행에서, 표본공간은 에너지 함수에 따라 10개의 소구역 $E_1 = \{\xi_n : U(\xi_n) \leq 125\}, E_2 = \{\xi_n : 125 < U(\xi_n) \leq 126\}, \cdots, E_{10} = \{\xi_n : U(\xi_n) > 135\}$으로 분할되었다. 그리고 여기에서 $U(\xi_n) = -\log \Pi(\xi_n | D^n)$은 모형 ξ_n의 음의 로그 사후확률이고 소위 말하는 에너지 함수이다.

그런 다음 10개의 데이터셋에 SVS를 적용했다. FDR 수준 0.0003에서 선택된 중요한 변수들은 SAMC에서 선택된 실제 중요한 변수들을 모두 포함했다. SVS의 실행을 측정하기 위해, false selection rate(FSR)와 negative selection rate(NSR)을 계산했다. s_i^*는 데이터셋 i를 위해 선택된 중요한 변수들의 집합을 나타내고, s는 실제로 중요한 변수들의 집합을 나타낸다.

$$FSR = \frac{\sum_{i=1}^{10} |s_i^* \setminus s|}{\sum_{i=1}^{10} |s_i^*|}, \qquad NSR = \frac{\sum_{i=1}^{10} |s \setminus s_i^*|}{\sum_{i=1}^{10} |s|}. \quad \cdots(35)$$

FSR과 NSR을 위와 같이 정의하고, 여기에서 $|\bullet|$은 집합의 원소 개수를 의

미한다. FSR과 NSR의 값이 작으면 작을수록 그 방법의 실행 효율은 더 좋다고 할 수 있다. 그 결과는 표 2에 요약되어 있다.

<표 2> 모의실험 데이터에서 각 방법에 대한 FSR, NSR 비교

	BSR ($\gamma = 0.85$)				
Methods	MAP	SVS	ridge	lasso	elastic net
Size	2.4	2.6	17	1	1
FSR(%)	0.48	0.46	2.4	0.8	0.8
NSR(%)	0.6154	0.5897	0.7059	4	4

능형, 라소, 엘라스틱넷의 벌점화 우도 방법들도 비교를 위해 데이터에 적용되었다. 라소와 엘라스틱넷 방법은 R의 'glmnet' 패키지를 사용했다. 라소와 엘라스틱넷에서의 벌점화 항은 $P_{\alpha,\lambda}(\beta) = \lambda[(1-\alpha)\frac{1}{2}|\beta_{[-0]}|_{l_2}^2 + \alpha|\beta_{[-0]}|_{l_1}]$이고, 여기에서 $0 < \alpha \leq 1$, $\beta_{[-0]}$은 절편을 제외한 회귀계수들의 벡터를 나타내며, $|\bullet|_{l_2}$그리고 $|\bullet|_{l_1}$은 각각 벡터의 L_2와 L_1 노음을 나타낸다. 모수 λ는 RLOOCV 오분류율을 최소화함으로써 선택된다.

SVS-BSR은 이 예시 데이터에 대해서 다른 방법들보다 현저하게 좋은 결과를 냈다. 능형, 라소, 엘라스틱넷과 비교하여 모든 10개의 데이터셋에 대하여 매우 낮은 FSR과 NSR에서 중요한 변수들을 선택하는 것이 가능했다. SAMC를 통해 선택된 MAP 모형들에 비해서도 SVS 과정을 거친 모형들의 FSR과 NSR이 비교적 낮은 것을 확인할 수 있다.

각각의 데이터셋에 대하여, 라소와 엘라스틱넷 둘 다 오직 1개의 중요한 변수만 선택했고, 이 변수는 상호작용 효과가 아닌 세 번째 변수의 세 번째 범주의 주효과이다. 표 2를 볼 때, MAP과 SVS의 크기는 절편항을 포함하지만 능형, 라소, 엘라스틱넷의 크기는 절편항을 포함하지 않는다. 다른 벌점화 우도 방법들과의 비교는 능형은 모든 변수들을 선택하고, 라소와 엘라스틱넷은 하나의 변수만 선택하여 정확한 선택을 하지 못한 것을 보여준다. 라소와 엘라스틱

넷의 NSR값이 높게 나타나는 것에서 볼 수 있듯, 실제로 유의한 변수들을 찾아내지 못하는 경향이 있다. MAP-BSR 모형과 SVS-BSR 모형은 모두 유의하게 생성한 두 가지 상호작용 효과를 포함했다. 이들은 벌점화 우도 방법들에 비해 중요한 변수들을 잘 선택하기 때문에 낮은 NSR값을 갖는다. FSR과 NSR을 모두 고려했을 때, 다른 모든 방법들에 비해 현저히 낮은 값을 갖는 것으로 보아 이 예시에서 가장 좋은 방법은 SVS-BSR 과정이라고 할 수 있다.

제 6 장 실증분석

BSR 사후 분포와 SVS 과정을 통하여 분할표에서의 고차 상호작용효과를 찾기 위해 제안된 방법의 효율성을 보이기 위해 네 가지 데이터를 이용하였다. SVS-BSR 방법은 R 패키지 'glmnet'을 이용하여 벌점화 우도 방법인 능형, 라소, 엘라스틱넷 방법과 비교한다. 각 방법은 RLOOCV를 계산하여 비교하였다. 비교한 표에서 rcvm은 최소 RLOOCV 값을 나타낸다. R 패키지 'glmnet'에서 $\alpha=0$은 능형, $\alpha=1$은 라소, $0<\alpha<1$은 엘라스틱넷을 실행한다. 본 논문에서는 $\alpha=0.5$를 사용하여 엘라스틱넷을 실행했다.

6.1 보리 가루 흰 곰팡이 균의 유전자 데이터

분석에 사용한 첫 번째 데이터는 보리 가루 흰 곰팡이 균의 유전자 데이터이다. 이 기생 곰팡이는 염색체 세트의 수가 절반으로 줄어든 반수체이기 때문에 유전학적으로 단순하다. 데이터에서 총 사례 수는 70건이고 분할표는 64개의 셀을 가진다. 편의상 6개의 유전자를 $A,B,\cdots,F$라고 두고, 해당 유전자를 가지면 1, 아니면 2로 표기한 이항 데이터이다.

<표 3> 보리 가루 흰 곰팡이 균의 유전자 데이터 분할표

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	3	0	1	0	
		2	0	1	0	0	0	1	0	0	
	2	1	1	0	1	0	7	1	4	0	
		2	0	0	0	2	1	3	0	11	
2	1	1	16	1	4	0	1	0	0	0	
		2	1	4	1	4	0	0	0	1	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

본 데이터는 표본이 64(= n)이고, 1개의 절편, 6개의 주효과, 15개의 1차 상호작용효과와 20개의 2차 상호작용효과, 그리고 15개의 3차 상호작용효과, 6개의 4차 상호작용효과를 포함한 63(= P_n)개의 예측치인 로그 선형데이터로 변환시켜 분석을 진행했다.

본 데이터를 위해 SAMC는 $t_0 = 100$, $K_n = 63$과 $\gamma = 0.85$로 선택하여 실행하였다. 각각의 값을 가지고 SAMC는 5.1×10^5의 반복을 했고, 여기서 처음 10,000번의 실행은 번인(burn-in)을 위해 버리고, 나머지만 추론을 위해 사용되었다. 표본공간은 $U(\xi_n)$에 따라 10개의 영역으로 분할하였다. $U(\xi_n) = -\log \Pi(\xi_n | D^n)$은 에너지 함수로 에너지 함수에 따라 표본공간이 다음과 같이 분할된다. 즉, $E_1 = \{\xi_n : U(\xi_n) \le 58\}$, $E_2 = \{\xi_n : 58 < U(\xi_n) \le 59\}$, $\cdots$, $E_{10} = \{\xi_n : U(\xi_n) > 68\}$ 이다.

분석 결과, MAP-BSR 모형은 $\{A, C, F, AB, AD, BD, CF\}$ 7개의 효과를 포함하고 있다. MAP-BSR 모형을 식으로 나타내면 다음과 같다.

$$\begin{aligned}\log(\mu_{ABCDEF}) = &-0.7419 + 2.23\lambda_A - 1.8971\lambda_C - 20.0322\lambda_{AB} - 2.7408\lambda_{AD} \\ &+ 2.0919\lambda_{BD} + 2.1972\lambda_{CF}.\end{aligned} \quad \cdots(36)$$

또한 SVS-BSR 모형은 $\{BEF, E, BCEF, BD, CF, F, AB, A, C, AD\}$와 같이 10개의 효과를 포함하고 있다. SVS-BSR 모형을 식으로 나타내면 다음과 같다.

$$\begin{aligned}\log(\mu_{ABCDEF}) = &\ 0.2362 - 17.2888\lambda_{BEF} - 0.8308\lambda_E + 19.3651\lambda_{BCEF} \\ &+ 1.8218\lambda_{BD} + 4.4849\lambda_{CF} - 2.8343\lambda_F - 21.9575\lambda_{AB} + 2.3985\lambda_A \\ &- 2.539\lambda_C - 2.7408\lambda_{AD}.\end{aligned} \quad \cdots(37)$$

비교를 위해 능형, 라소, 엘라스틱넷 방법을 곰팡이 균의 유전자 데이터에 적용해보았다. 표 4에 나타난 비교 결과를 보면, 라소는 1.9123의 rcvm를 가지는 LSI를 제외한 변수 19개, 엘라스틱넷은 2.2288의 rcvm를 가지는 LSI를 제외한

변수 41개, 능형은 2.4938의 rcvm를 갖는 모든 변수 62개를 선택한다. BSR은 MAP 모형이 11개의 변수를 선택하지만 rcvm는 1.2054로 매우 작다. SVS 방법을 거친 BSR은 1.3847의 rcvm를 갖는 10개의 변수를 선택했고, 변수를 선택하는 q-value 기준을 늘려가며 살펴본 11개 변수의 rcvm는 1.4253, 그리고 12개 변수의 rcvm는 1.6737이다. 따라서 BSR이 변수선택에 있어서 기존의 벌점화우도 방법들보다 훨씬 우수한 것을 보여준다.

<표 4> 보리 가루 흰 곰팡이 균의 유전자 데이터에 대한 변수선택 방법의 변수의 개수와 rcvm 비교

	BSR ($\gamma=0.85$)									
	MAP		SVS		ridge		lasso		elastic net	
Data	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm
Gene	11	1.2054	10	1.3847	62	2.4638	19	1.9123	41	2.2288

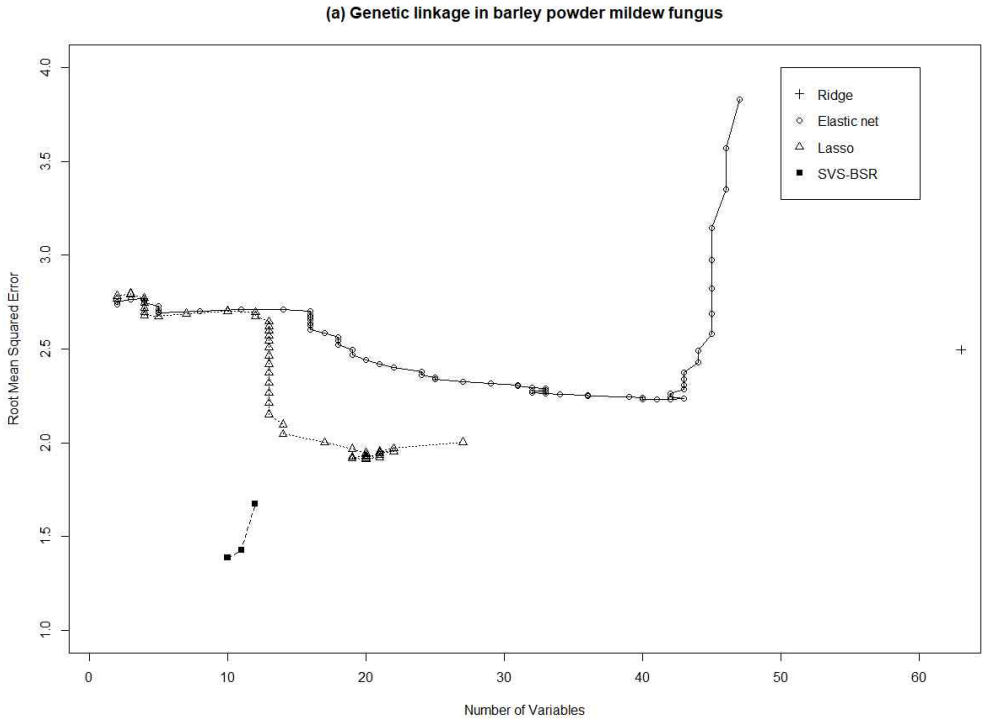
<표 5> SVS-BSR에서 변수 개수에 따른 rcvm

size	10	11	12
rcvm	1.3847	1.4253	1.6737

정규화 경로(regularization path)에 따라 능형, 라소, 엘라스틱넷에 의해 선택된 모형의 rcvm과 BSR과 SVS에 의해 추출된 최종 모형의 최소 rcvm을 비교하였다. 그림 1에서 볼 수 있듯 SVS-BSR 방법은 다른 방법들에 비해서 변수를 적게 선택하지만 훨씬 작은 rcvm을 가지는 것으로 보아 적은 수의 실제로 중요한 변수들 모형으로 가장 효율적인 모형이다.

전진 변수선택 성질 때문에 라소와 엘라스틱넷은 노이즈 변수들에 민감하고 P_n 차원에 의해 불리한 영향을 받는 경향이 있다. 그러나 SVS-BSR 방법은 전체 모형공간에서 포괄적인 탐색을 하고 모든 변수들의 결합정보를 사용한다. 이런 이유로 SVS-BSR은 노이즈 변수들에 민감하지 않고 P_n이 증가함에 영향을 덜 받는다. 이 비교는 SVS-BSR이 벌점화 우도 방법들보다 고차원 변수선

택에 더 적절하다는 것을 보여준다.



<그림 1> 보리 가루 흰 곰팡이 균의 유전자 데이터에 대한 변수 선택 방법의 변수의 개수와 rcvm 비교

6.2 체코 자동차 공장 근로자 데이터

다음으로 체코의 자동차 공장 근로자들에 대한 데이터를 다룬다. 데이터는 15년 후속 연구의 시작에서 수집된 체코 슬로바키아 자동차 공장의 근로자들의 관상동맥 혈전증의 잠재적 위험요소에 대한 조사의 일부이다. 여섯 가지의 가능한 위험요소를 $A, B, \cdots, F$라고 둔 이항 데이터이다. 본 데이터는 표본이 64 $(=n)$이고, 1개의 절편, 6개의 주효과, 15개의 1차 상호작용효과와 20개의 2차 상호작용효과, 그리고 15개의 3차 상호작용효과, 6개의 4차 상호작용효과를 포

함한 63($=P_n$) 개의 예측치인 로그 선형데이터로 변환시켜 분석을 진행했다.

<표 6> 체코 슬로바키아 자동차 공장의 근로자 데이터 분할표

			1				2				C
			1		2		1		2		B
			1	2	1	2	1	2	1	2	A
1	1	1	44	40	112	67	129	145	12	23	
		2	35	12	80	33	109	67	7	9	
	2	1	23	32	70	66	50	80	7	13	
		2	24	25	73	57	51	63	7	16	
2	1	1	5	7	21	9	9	17	1	4	
		2	4	3	11	8	14	17	5	2	
	2	1	7	3	14	14	9	16	2	3	
		2	4	0	13	11	5	14	4	4	
F	E	D									

본 데이터를 위해 SAMC는 $t_0=100$, $K_n=63$과 $\gamma=0.85$로 정하고 5.1×10^5의 반복을 했고, 여기서 처음 10,000번의 실행은 번인을 위해 버리고, 나머지만 추론을 위해 사용되었다. 표본공간은 에너지 함수에 따라 $U(\xi_n)$에 따라 10개의 영역으로 분할하였다. $U(\xi_n)=-\log\Pi(\xi_n|D^n)$은 에너지 함수로 $E_1=\{\xi_n: U(\xi_n)\le 110\}$, $E_2=\{\xi_n: 110<U(\xi_n)\le 111\}$, $\cdots$, $E_{10}=\{\xi_n: U(\xi_n)>120\}$ 이다. MAP-BSR 모형은 $\{B,C,E,F,AC,AD,BC,BE,ADE\}$와 같이 9개의 효과를 포함하고 있다. MAP-BSR 모형을 식으로 나타내면 다음과 같다.

$$\begin{aligned}\log(\mu_{ABCDEF}) = {} & 3.73002+0.72035\lambda_B+0.89779\lambda_C-0.61617\lambda_E \\ & -1.80513\lambda_F+0.38661\lambda_{AC}-0.82768\lambda_{AD}-2.8015\lambda_{BC} \\ & +0.41897\lambda_{BE}+0.6777\lambda_{ADE}.\end{aligned} \quad \cdots(38)$$

SVS-BSR 모형은 $\{BCDF,BE,CE,E,AC,ADE,B,BC,C,F,AD\}$ 11개의 효과를 포함하고 있다. SVS-BSR 모형을 식으로 나타내면 다음과 같다.

$$\log(\mu_{ABCDEF}) = 3.65909 + 0.82826\lambda_{BCDF} + 0.26292\lambda_{BE} - 0.26149\lambda_{CE} - 0.42328\lambda_{E} + 0.39878\lambda_{AC} + 0.68828\lambda_{ADE} + 0.77972\lambda_{B} - 2.85747\lambda_{BC} + 0.99469\lambda_{C} - 1.83816\lambda_{F} - 0.84402\lambda_{AD}. \quad \cdots(39)$$

비교를 위해 능형, 라소, 엘라스틱넷 방법을 데이터에 적용해보았다. 표 6에 나타난 비교 결과를 보면, 라소는 8.8236의 rcvm를 가지는 LSI를 제외한 변수 41개, 엘라스틱넷은 9.8332의 rcvm를 가지는 LSI를 제외한 변수 43개, 능형은 17.9588의 rcvm를 갖는 모든 변수 62개를 선택한다. BSR은 MAP 모형이 17개의 변수를 선택하지만 rcvm는 5.2897로 매우 작다. SVS 방법을 거친 BSR은 8.5465의 rcvm를 갖는 16개의 변수를 선택했고, 변수를 선택하는 q-value 기준을 늘려가며 살펴본 12개 변수의 rcvm는 8.9358, 13개 변수의 rcvm는 8.7815, 14개 변수의 rcvm는 8.6328, 그리고 15개 변수의 rcvm는 8.7386이다. 따라서 BSR이 변수 선택에 있어서 기존의 벌점화우도 방법들보다 훨씬 우수한 것을 보여준다.

<표 7> 체코 슬로바키아 자동차 공장의 근로자 데이터에 대한 변수 선택 방법의 변수의 개수와 rcvm 비교

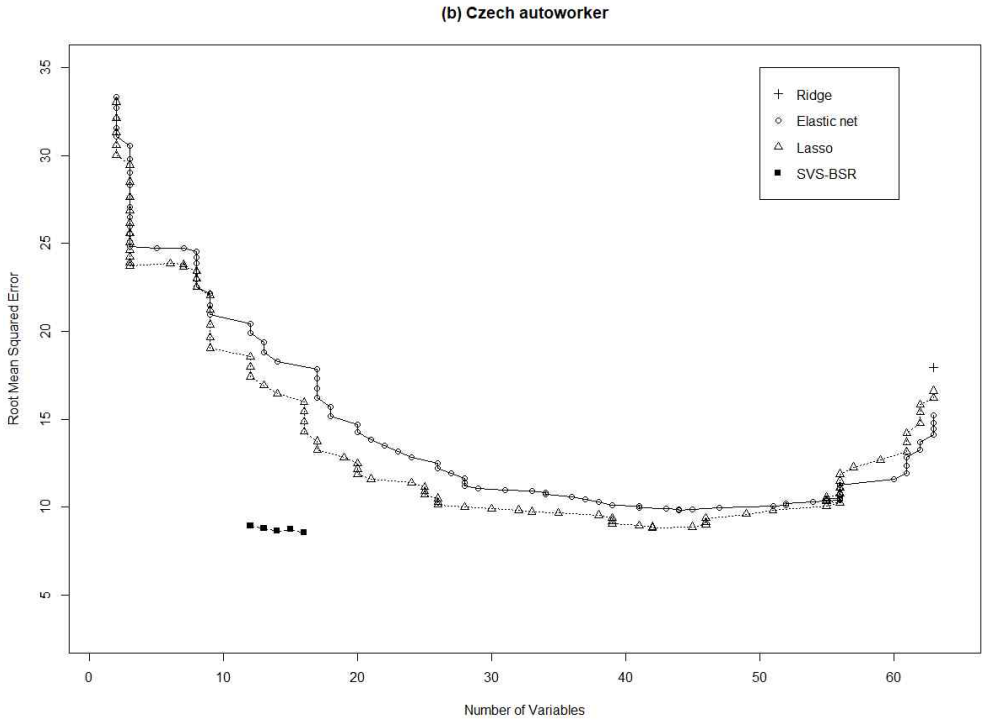
	BSR ($\gamma = 0.85$)									
	MAP		SVS		ridge		lasso		elastic net	
Data	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm
Workers	17	5.2897	16	8.5465	62	17.9588	41	8.8236	43	9.8332

<표 8> SVS-BSR에서 변수 개수에 따른 rcvm

size	12	13	14	15	16
rcvm	8.9358	8.7815	8.6328	8.7386	8.5465

정규화 경로에 따라 능형, 라소, 엘라스틱넷에 의해 선택된 모형의 rcvm과 MAP-BSR과 SVS-BSR에 의해 추출된 최종 모형의 최소 rcvm을 비교하였다. 그림 2에서 볼 수 있듯 SVS-BSR 모형이 다른 모형들보다 rcvm이 훨씬 작아

좋은 결과를 냈다고 할 수 있다.



<그림 2> 체코 슬로바키아 자동차 공장의 근로자 데이터에 대한 변수선택 방법의 변수의 개수와 rcvm 비교

6.3 로치데일 가정 조사 데이터

다음으로 로치데일 지역의 가정 조사 데이터를 다룬다. 데이터는 로치데일 가정 조사에서 여성의 경제활동 및 남편의 실업과 관련된 8가지 이항 변수의 교차 분류이다. A는 아내가 경제활동을 하는지 여부, B는 아내의 나이가 38세 이상인지 여부, C는 남편이 실직했는지 여부, D는 자식이 4살 이하인지 여부, E는 아내의 학력이 고등학교 이상인지 여부, F는 남편의 학력이 고등학교 이상인지 여부, G는 아시아 출신인지 여부, H는 일하는 다른 가구구성원의 여부

를 나타낸다. 분할표에서 256개의 셀에 분류된 665명의 개체가 있다. 본 데이터는 표본이 256(=n)이고, 1개의 절편, 8개의 주 효과, 28개의 1차 상호작용효과와 56개의 2차 상호작용효과, 그리고 70개의 3차 상호작용효과, 56개의 4차 상호작용효과, 28개의 5차 상호작용효과, 8개의 6차 상호작용효과를 포함한 255(=P_n) 개의 예측치인 로그 선형데이터로 변환시켜 분석을 진행했다.

<표 9> 로치데일 지역의 가정 조사 데이터 분할표

				Y								N								H
				Y				N				Y				N				G
				Y		N		Y		N		Y		N		Y		N		F
				Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	E
Y	Y	Y	Y	5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0	
			N	8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0	
		N	Y	5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	
			N	4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0	
	N	Y	Y	17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0	
			N	1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	
		N	Y	4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0	
			N	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	
N	Y	Y	Y	18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0	
			N	5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1	
		N	Y	3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0	
			N	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	N	Y	Y	41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0	
			N	0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0	
		N	Y	2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0	
			N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A	B	C	D																	

본 데이터를 위해 SAMC는 $t_0 = 100$, $K_n = 255$과 $\gamma = 0.85$로 정하고 5.1×10^5의 반복을 했고, 여기서 처음 10,000번의 실행은 번인을 위해 버리고, 나머지만 추론을 위해 사용되었다. 표본공간은 에너지 함수 $U(\xi_n)$에 따라 10개의 영역으로 분할하였다. $U(\xi_n) = -\log \Pi(\xi_n | D^n)$은 에너지 함수로 $E_1 = \{\xi_n : U(\xi_n) \le 272\}$, $E_2 = \{\xi_n : 272 < U(\xi_n) \le 273\}$, $\cdots$, $E_{10} = \{\xi_n : U(\xi_n) > 282\}$ 이다. MAP-BSR 모형은 $\{A, B, E, F, AE, AF, BD, CE, CF, EF, ABF, ACD, ACF, BCF, BDF\}$ 와 같이 15개의

효과를 포함하고 있다. MAP-BSR 모형을 식으로 나타내면 다음과 같다.

$$\begin{aligned}\log(\mu_{ABCDEFGH}) = {} & 1.51483 + 1.01505\lambda_A + 0.56127\lambda_B - 0.87024\lambda_E \\ & - 1.43666\lambda_F + 0.16019\lambda_{AE} - 1.81836\lambda_{AF} - 3.42091\lambda_{BD} \\ & - 1.69813\lambda_{CE} + 0.57066\lambda_{CF} - 0.52561\lambda_{EF} - 1.73264\lambda_{ABF} \\ & - 3.36489\lambda_{ACD} - 14.91606\lambda_{ACF} - 0.69857\lambda_{BCF} + 2.185\lambda_{BDF}.\end{aligned} \quad \cdots(40)$$

또한 SVS-BSR 모형은 $\{CD, BDE, ABCDEF, BCDFGH, ABF, BCF, CF, BD, EF, ACD, AE, E, AF, BDF, B, CE, A, F, ACF\}$와 같이 19개의 효과를 포함하고 있다. SVS-BSR 모형을 식으로 나타내면 다음과 같다.

$$\begin{aligned}\log(\mu_{ABCDEFGH}) = {} & 1.53659 - 0.18225\lambda_{CD} - 15.6988\lambda_{BDE} \\ & + 22.27788\lambda_{ABCEF} - 14.61242\lambda_{BCDFGH} - 1.73651\lambda_{ABF} \\ & - 0.73929\lambda_{BCF} + 0.68644\lambda_{CF} - 3.06047\lambda_{BD} \\ & - 0.40052\lambda_{EF} - 3.20228\lambda_{ACD} + 0.16772\lambda_{AE} \\ & - 0.86227\lambda_E - 1.79729\lambda_{AF} + 2.17578\lambda_{BDF} \\ & + 0.55068\lambda_B - 1.69812\lambda_{CE} + 0.99551\lambda_A \\ & - 1.49337\lambda_F - 15.44514_{ACF.}\end{aligned} \quad \cdots(41)$$

비교를 위해 능형, 라소, 엘라스틱넷 방법을 데이터에 적용해보았다. 표 8에 나타난 비교 결과를 보면, 라소는 2.778의 rcvm를 가지는 LSI를 제외한 변수 45개, 엘라스틱넷은 3.0268의 rcvm를 가지는 LSI를 제외한 변수 53개, 능형은 4.1074의 rcvm를 갖는 모든 변수 254개를 선택한다. BSR은 MAP 모형이 26개의 변수를 선택하지만 rcvm는 1.6748로 매우 작다. SVS 방법을 거친 BSR은 1.7313의 rcvm를 갖는 23개의 변수를 선택했고, 변수를 선택하는 q-value 기준을 늘려가며 살펴본 20개 변수의 rcvm는 1.8936, 21개 변수의 rcvm는 1.8992, 22개 변수의 rcvm는 1.7331, 24개 변수의 rcvm는 1.8035, 25개 변수의 rcvm는 1.8245, 26개 변수의 rcvm는 1.8317, 27개 변수의 rcvm는 1.8360, 28개 변수의 rcvm는 1.8148, 29개 변수의 rcvm는 1.8914, 그리고 30개 변수의 rcvm는 1.8906이다. 따라서 SVS-BSR이 변수선택에 있어서 기존의 벌점화우도 방법들

보다 훨씬 우수한 것을 보여준다.

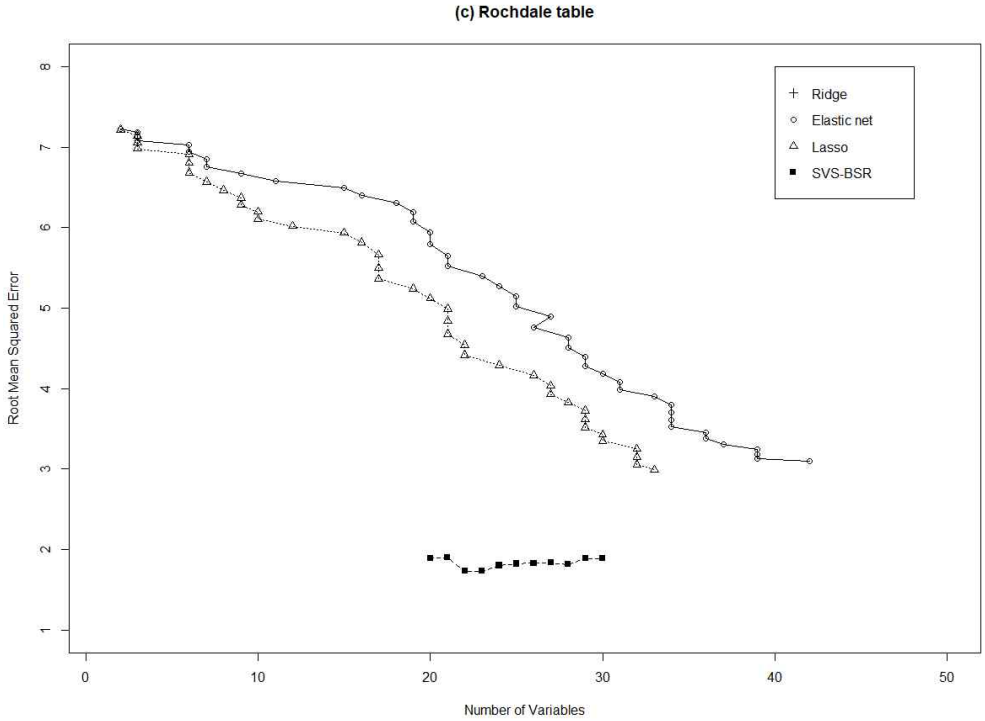
<표 10> 로치데일 지역의 가정 조사 데이터에 대한 변수 선택 방법의 변수의 개수와 rcvm 비교

	BSR ($\gamma = 0.85$)									
	MAP		SVS		ridge		lasso		elastic net	
Data	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm
Household	26	1.6748	23	1.7313	254	4.1074	45	2.778	53	3.0268

<표 11> SVS-BSR에서 변수 개수에 따른 rcvm

size	20	21	22	23	24	25	26	27	28	29	30
rcvm	1.8936	1.8992	1.7331	1.7313	1.8035	1.8245	1.8317	1.836	1.8148	1.8914	1.8906

정규화 경로에 따라 능형, 라소, 엘라스틱넷에 의해 선택된 모형의 rcvm과 SVS-BSR에 의해 추출된 최종 모형의 최소 rcvm을 비교하였다. 그림 3에서 볼 수 있듯, SVS-BSR 모형이 다른 모형들보다 rcvm이 훨씬 작아 좋은 결과를 냈다고 할 수 있다.



<그림 3> 로치데일 지역의 가정 조사 데이터에 대한 변수 선택 방법의 변수의 개수와 rcvm 비교

6.4 소득별 종합 출근 데이터

다음으로 소득별 종합 출근 데이터를 다룬다. 각각 4개의 집 영역, 4개의 직장 영역, 그리고 16개의 소득 범주를 갖는 범주형 데이터이다. 셀에 0이 많아 데이터의 희소성을 보이는 것은 일부 지역에는 저소득층 근로자가 없기 때문이다. 본 데이터는 표본이 256($=n$)이고, 1개의 절편, 21개의 주 효과, 99개의 1차 상호작용효과를 포함한 121($=P_n$) 개의 예측치인 로그 선형데이터로 변환시켜 분석을 진행했다.

<표 12> 소득별 종합 출근 데이터 분할표

Home Zone	Work Zone	Income Category																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	C
a	a	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	b	46	34	0	23	0	0	0	0	0	0	0	0	0	0	0	0	
	c	243	200	0	0	45	0	0	0	70	0	0	80	0	0	0	0	
	d	0	0	0	0	0	0	0	45	60	0	0	0	0	0	0	0	
b	a	4	9	15	14	18	17	0	0	17	18	22	44	33	0	16	16	
	b	0	0	0	0	0	0	0	0	0	0	0	0	0	78	0	0	
	c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
c	a	14	24	36	34	14	16	17	18	0	18	12	0	44	34	33	33	
	b	0	0	14	0	16	18	18	34	12	16	44	22	16	18	12	14	
	c	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	
	d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
d	a	12	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	b	14	12	67	9	22	66	14	14	34	37	38	12	24	22	16	18	
	c	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	
	d	0	0	0	0	0	0	18	0	0	22	0	0	0	0	0	0	
A	B																	

본 데이터를 위해 SAMC는 $t_0 = 100$, $K_n = 255$과 $\gamma = 0.85$로 정하고 5.1×10^5의 반복을 했고, 여기서 처음 10,000번의 실행은 번인을 위해 버리고, 나머지만 추론을 위해 사용되었다. 표본공간은 에너지 함수 $U(\xi_n)$에 따라 10개의 영역으로 분할하였다. $U(\xi_n) = -\log \Pi(\xi_n | D^n)$은 에너지 함수로 $E_1 = \{\xi_n : U(\xi_n) \le 10\}$, $E_2 = \{\xi_n : 10 < U(\xi_n) \le 11\}$, $\cdots$, $E_{10} = \{\xi_n : U(\xi_n) > 20\}$ 이다. MAP-BSR 모형은 $\{C_{16}\}$을 포함하고 있다. MAP-BSR 모형을 식으로 나타내면 다음과 같다.

$$\log(\mu_{ABC}) = 2.11726 + 0.80388\lambda_{C_{16}}. \quad \cdots(42)$$

SVS-BSR 모형은 $\{C_{15}, B_b, C_{16}, C_4, A_d C_6, A_d C_4, A_b C_{12}, A_b C_{15}, A_d C_{14}, C_6, A_d B_d, B_b C_{14}, B_c C_6, A_d C_{16}, B_b C_{11}, B_c\}$와 같이 16개 효과를 포함하고 있다. SVS-BSR 모형을 식

으로 나타내면 다음과 같다.

$$\begin{aligned}\log(\mu_{ABC}) = {} & 2.49255 - 0.25197\lambda_{c_{15}} - 2.88606\lambda_{B_b} \\ & + 0.05743\lambda_{c_{16}} - 0.21201\lambda_{c_4} + 0.32592\lambda_{A_dC_6} \\ & + 1.44886\lambda_{A_dC_4} + 0.7722\lambda_{A_bC_{12}} + 1.03921\lambda_{A_bC_{15}} \\ & + 3.10156\lambda_{A_dC_{14}} - 0.22969\lambda_{C_6} - 1.04722\lambda_{A_dB_d} \\ & + 1.13356\lambda_{B_bC_{14}} + 2.12595\lambda_{B_cC_6} + 2.81388\lambda_{A_dC_{16}} \\ & + 1.40481\lambda_{B_bC_{11}} - 0.7639\lambda_{B_c}.\end{aligned} \quad \cdots(43)$$

비교를 위해 능형, 라소, 엘라스틱넷 방법을 데이터에 적용해보았다. 표 10에 나타난 비교 결과를 보면, 라소는 23.8852의 rcvm를 가지는 LSI를 제외한 변수 26개, 엘라스틱넷은 24.0384의 rcvm를 가지는 LSI를 제외한 변수 39개, 능형은 24.0165의 rcvm를 갖는 변수 120개를 선택한다. BSR은 MAP 모형이 6개의 변수를 선택하지만 rcvm는 1.0174로 매우 작다. SVS 방법을 거친 BSR은 0.8726의 rcvm를 갖는 18개의 변수를 선택했고, 변수를 선택하는 q-value 기준을 늘려가며 살펴본 14개 변수의 rcvm는 0.9488, 15개 변수의 rcvm는 0.944, 16개 변수의 rcvm는 0.9103, 그리고 17개 변수의 rcvm는 0.9128이다. 따라서 BSR이 변수 선택에 있어서 기존의 벌점화우도 방법들보다 훨씬 우수한 것을 보여준다.

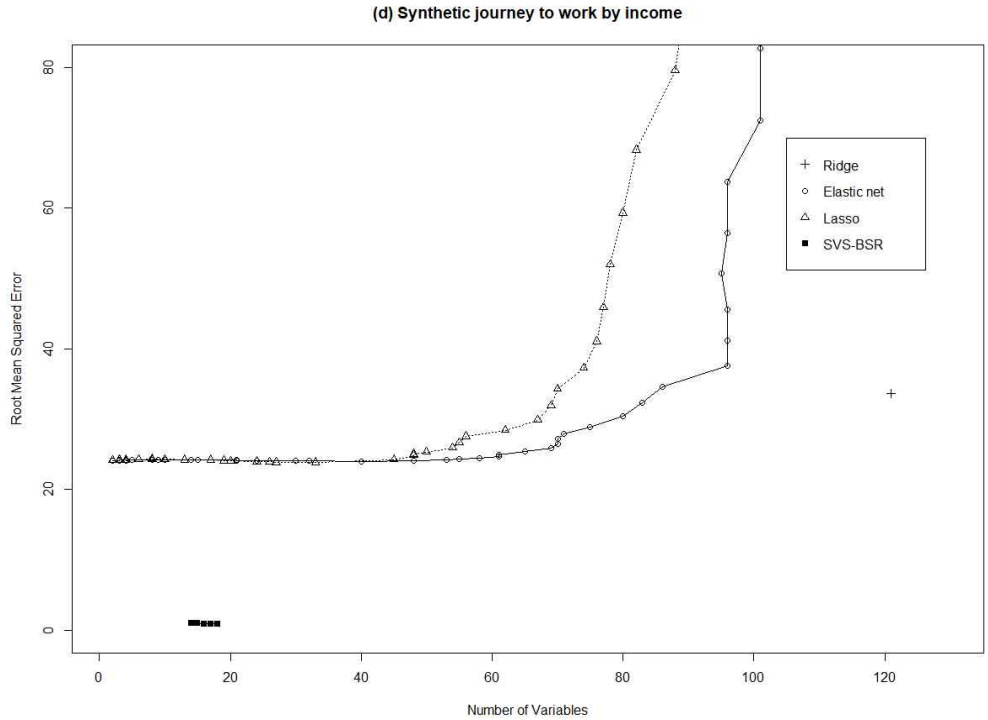
<표 13> 소득별 종합 출근 데이터에 대한 변수선택 방법의 변수의 개수와 rcvm 비교

	BSR ($\gamma = 0.85$)									
	MAP		SVS		ridge		lasso		elastic net	
Data	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm	$\lvert\xi_n\rvert$	rcvm
Workers	6	1.0174	18	0.8726	120	24.0165	26	23.8852	39	24.0384

<표 14> SVS-BSR에서 변수 개수에 따른 rcvm

size	14	15	16	17	18
rcvm	0.9489	0.944	0.9103	0.9128	0.8726

정규화 경로에 따라 능형, 라소, 엘라스틱넷에 의해 선택된 모형의 rcvm과 SVS-BSR에 의해 추출된 최종 모형의 최소 rcvm을 비교하였다. 그림 4에서 볼 수 있듯, SVS-BSR 모형이 다른 모형들보다 rcvm이 훨씬 작아 좋은 결과를 냈다고 할 수 있다.



<그림 4> 소득별 종합 출근 데이터에 대한 변수선택 방법의 변수의 개수와 rcvm 비교

오직 BSR로만 추출된 MAP 모형들은 변수를 너무 적게 선택하여, 중요한 변

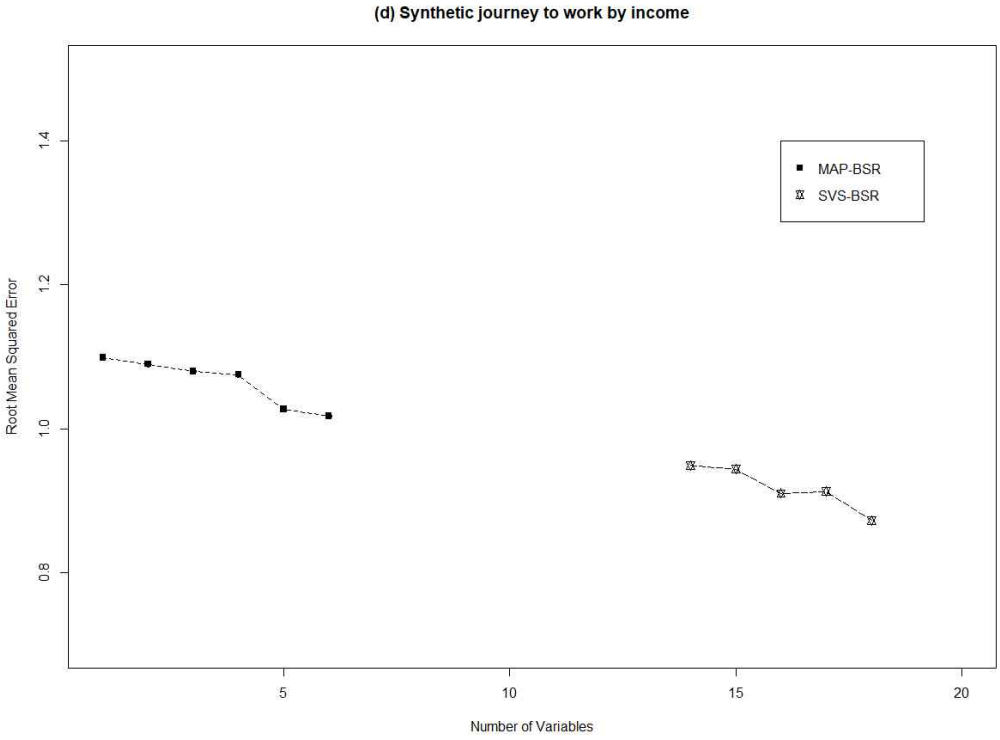
수들을 놓칠 수도 있다. 하지만 SVS-BSR 모형들은 MAP-BSR 모형들보다 변수들을 더 많이 선택하지만 rcvm 값이 현저히 낮은 것을 확인할 수 있다. 이것은 SVS 과정이 MAP 모형이 놓친 실제로 중요한 변수들을 찾아내어 모형에 포함시켰다는 것을 보여준다.

<표 15> MAP-BSR 과정에서 변수의 개수에 따른 rcvm

size	1	2	3	4	5	6
rcvm	1.09865	1.08972	1.07952	1.07529	1.02708	1.01742

<표 16> SVS-BSR 과정에서 변수의 개수에 따른 rcvm

size	14	15	16	17	18
rcvm	0.948849	0.944005	0.910295	0.912791	0.872571



<그림 5> 소득별 종합 출근 데이터에 대한 MAP-BSR 방법과 SVS-BSR 방법의 변수의 개수와 rcvm 비교

제 7 장 결 론

본 연구에서는 여러 개의 범주형 변수들의 교차 분류된 포아송 분포 분할표의 변수선택을 위해 SVS-BSR 방법을 제안했다. 변수의 수가 증가하거나 범주의 수가 증가하면 데이터의 차원이 커져 정확한 분석이 어려워지기 때문이다. BSR 사후분포로부터 표본추출을 효율적으로 하기 위해 SAMC 알고리즘을 적용했다. 더 나아가 선택된 변수들이 실제로 유의하게 중요한 변수들인지 확인하기 위해 FDR을 이용하여 SVS 검사를 거쳤다.

제안된 변수 검사 과정을 포함한 SVS-BSR 방법의 우수성을 보기 위해, 고차원 분할표 데이터에 대해 SVS-BSR 방법을 기존의 벌점화우도 방법들인 능형, 라소, 엘라스틱넷 방법들과 비교를 진행했다. 비교 결과, 모든 분석에서 SVS-BSR에 의해 선택된 모형들이 변수의 수가 다른 방법들에 비해 작고, RLOOCV 값도 가장 작아 SVS-BSR 방법이 벌점화우도 방법들보다 우수한 것을 알 수 있었다.

BSR을 통해 만들어진 MAP-BSR 모형과 SVS 절차를 거친 BSR을 통해 만들어진 SVS-BSR 모형 비교에서 MAP-BSR 모형들은 변수를 적게 선택하지만 중요한 변수를 놓칠 수 있다는 것을 보였다. 그에 반해 SVS-BSR 모형들은 MAP 모형보다는 변수들을 더 많이 선택하지만 rcvm 값이 낮아 MAP 모형이 놓칠 수도 있는 실제로 중요한 변수들을 포함했다. SVS 과정은 고차원 데이터의 차원이 커질수록 더 큰 강점을 보이는 것 또한 확인할 수 있었다.

참 고 문 헌

[1] Agresti, A. (2007). *An introduction to categorical data analysis*, 2nd ed., John Wiley & Sons, New York.

[2] Akdeniz, F. (2002). More on the Pre-test Estimator in Ridge Regression, *Communication in Statistics Theory and Method,* 31(6), 987-994.

[3] Baglivo, J., Olivier, D., Pagano, M. (1988). Methods for the Analysis of Contingency Tables with Large and Small Cell Counts, *Journal of the American Statistical Association,* 83(404), 1006-1013.

[4] Batah, F. M., Ramanathan, V., Gore, S. D. (2008). The efficiency of Modified Jackknife and Ridge Type Regression Estimators: A comparison, *Surveys in Mathematics and its Applications,* 24(2), 157-174.

[5] Benjamini, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society Series B,* 57(1), 289-300.

[6] Bottolo, L., Richardson, S. (2010), Evolutionary stochastic search for Bayesian model exploration, *International Society for Bayesian Analysis,* 5(3), 583-618.

[7] Cheon, S., Liang, F., Chen, Y., Yu, K. (2014). Stochastic approximation Monte Carlo importance sampling for approximating exact conditional probabilities, *Statistics and Computing,* 24(4), 505-520.

[8] Cheon, S. (2017). Analysis of High-Dimensional Contingency Table Using Bayesian Subset Regression, *Journal of the Korean Data Analysis Society,* 19(4B), 1841-1852. (in Korean)

[9] Dorugade, A. V., Kashid, D. N. (2010). Variable Selection in Linear regression Based on Ridge Estimator, *Journal of Statistical Association,* 38(3), 248-250.

[10] Foulkes, A. S., Reilly, M., Zhou, L., Wolfe, M. and Rader, D. J. (2005). Mixed modelling to characterize genotype-phenotype associations, *Statistical in Medicine*, 24(5), 775-789.

[11] Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software,* 33(1), 1-22.

[12] Golub, G. H., Health, M., Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, *Journal of Technometrices,* 21(2), 385-405.

[13] Haario, H., Saksman, E., Tamminen, J. (2001). An adaptive Metropolis algorithm, *Bernoulli Society for Mathematical Statistics and Probability,* 7(2), 223-242.

[14] Hans, C., Dobra, A., West, M. (2007). Shotgun Stochastic Search for "Large p" Regression, *Journal of the American Statistical Association,* 102(478), 507-516.

[15] Hoerl, A. E., Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Journal of Technometrices,* 12(1), 55-67.

[16] Hsu, J. S. J. (1995). Generalized Laplacian approximations in Bayesian inference, *The Canadian Journal of Statistics,* 23(4), 399-410.

[17] Hu, J., Joshi A., Johnson V. E. (2009). Log-Linear Models for Gene Association, *Journal of the American Statistical Association,* 104(486), 597-607.

[18] Jo, A., Cheon, S. (2015). The study of variable selection methods using the penalized likelihood methods in high-dimensional linear model, *Journal of the Korean Data Analysis Society*, 17(5B), 2391-2402. (in Korean)

[19] Jung, B. C., So, S., Cheon, S. (2014). Exact inference in contingency tables via stochastic approximation Monte Carlo, *Journal of the Korean Statistical Society,* 43, 31-45.

[20] Kass, R. E., Raftery, A. E. (1995). Bayes Factors, *Journal of the American Statistical Association,* 90(403), 773-795.

[21] Kuo, L., Mallick, B. (1998). Variable Selection for Regression Models, *Sankhyā: The Indian Journal of Statistics Series B*, 60, 65-81.

[22] Lee, S., Kwon, S. (2015). Moderately clipped LASSO for the sparse high-dimensional logistic regression models, *Journal of the Korean Data Analysis Society*, 17(3A), 145-154.

(in Korean)

[23] Liang, F., Wong, W. H. (2001). Real-Parameter Evolutionary Monte Carlo With Applications to Bayesian Mixture Models, *Journal of the American Statistical Association,* 96(454), 653-666.

[24] Liang, F., Liu, C., Carroll R. J. (2007). Stochastic Approximation in Monte Carlo Computation, *Journal of the American Statistical Association,* 102(477), 305-320.

[25] Liang, F., Song, Q., Yu, K. (2013). Bayesian Subset Modeling for High-Dimensional Generalized Linear Models, *Journal of the American Statistical Association,* 108(502), 589-606.

[26] Liang, F., Zahng, J. (2008). Estimating the false discovery rate using the stochastic approximation algorithm, Biometrika, 95(4), 961-977.

[27] McCullaph, P., Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall.

[28] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. (1953), Equation of State Calculations by Fast Computing Machines, *The Journal of Chemical Physics,* 21(6), 1087-1092.

[29] Nelder, J. A., Wedderburn, R. W. M., (1972). Generalized Linear Models, *Journal of the Royal Statistical Society Series A*, 135(3), 370-384.

[30] Oh, E. J., Le, H. (2013). Dimension reduction and prediction for high-dimensional regression models using the graphical lasso, *Journal of the Korean Data Analysis Society,* 15(5), 2321-232. (in Korean).

[31] Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society,* 64(3), 479-498.

[32] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value, *The Annals of Statistics,* 31, 2013-2035.

[33] Sun, X., Choi, H., Kwon, S. (2007). A sparse ridge estimation for the sparse logistic regression model, *Journal of the Korean Data Analysis Society,* 16(4A), 1715-1725. (in Korean).

[34] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society,* 58(1), 267-288.

[35] Wang, F., Landau, D. P. (2001). Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States, *American Physical Society*, 86(10), 2050-2053.

[36] Wu, M., Liang, F. (2011). Population SAMC vs SAMC: Convergence and Applications to Gene Selection Problems, *Journal of Biometrics & Biostatistics*, S1:002.

[37] Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society Series B,* 57(1), 289-300.

[38] Zou, H., Hastie, T. (2005). Addendum: Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society,* 67(5), 768-768