Structured Variational Inference for Bayesian Variable Selection
via Nested Sequential Monte Carlo Sampler

1. 선형회귀모형(linear regression)의 변수선택법(variable selection)

다음과 같이 연속형 반응변수 Y와 p개의 설명변수를 갖는 선형회귀모형을 고려하자.

$$Y = \alpha + \sum_{j=1}^{p} X_j \theta_j + \epsilon \tag{1}$$

α는 절편, 각 X_j는 j번째 설명변수, θ_j는 X_j에 해당하는 회귀계수, 그리고 ϵ은 정규분포 $N(0,\sigma^2)$를 따르는 오차이다. 변수선택법은 주어진 설명변수들 중에서 Y를 가장 잘 예측하는 최소한의 설명변수 집합을 찾는 것이 목표이고 이에 대한 방법론으로 후방 소거법(backward elimination), 단계적 선택법(stepwise selection), 축소추정 기법(shrinkage method) 등이 존재한다.

1.1 베이지안(bayesian) 변수선택법

베이지안 변수선택법은 각 설명변수 X_j가 회귀모형에 포함될 사후확률(posterior probability)을 계산하는 과정으로 구성된다. 이를 위해 각 회귀계수 θ_j은 다음과 같은 Spike and Slab 사전분포(prior distribution)를 부여할 수 있다(Mitchell and Beauchamp, 1988).

$$\begin{aligned} \boldsymbol{\theta} \,|\, \boldsymbol{\gamma}, \sigma^2 \;&\sim\; \prod_{j=1}^{p} \gamma_j \pi(\widetilde{\theta}_j | \sigma^2) + (1-\gamma_j)\delta_{0,} \\ \boldsymbol{\gamma} \,|\, \rho \;&\sim\; \prod_{j=1}^{p} Bernoulli(\rho) \end{aligned} \tag{2}$$

$\boldsymbol{\theta} = (\theta_1,...,\theta_p)^\top$은 p차원의 회귀계수 벡터, $\boldsymbol{\gamma} = (\gamma_1,...,\gamma_p)^\top$은 p차원의 이진(binary) 변수들 벡터, δ_0은 0에 점질량(point mass)을 가지는 비생성(degenerate) 분포, 그리고 ρ은 γ_j가 1이 될 확률을 나타내는 변수이다. 각 Slab 분포 $\pi(\widetilde{\theta}_j)$를 위한 사전분포로 정규분포를 고려할 수 있지만 과한 축소를 일으킬 여지가 있어 정규분포보다 두꺼운 꼬리를 가지는 라플라스 분포 $Laplace(0,\lambda/\sigma)$를 선택하였고 희소(sparse)한 변수선택을 위해 ρ의 사전분포로 베타분포 $Beta(1,p)$을 선택하였다(Castillo et al., 2015, Ray and Szabó, 2022). 절편 α과 오차의 분산 σ^2은 모든 가능한 모형들이 공통적으로 가지는 변수들이기 때문에 Jeffreys 사전 분포 $\pi(\alpha,\sigma^2) \propto 1/\sigma^2$를 부여하였다.

1.2. 보조변수(auxiliary variable)를 활용한 라플라스 분포 표현

2장에서 서술할 변분추론을 하기 위해 완전조건사후분포(fully conditional posterior distribution)을 구할 수 있어야 한다. 그러므로 라플라스 사전분포는 다음과 같이 지수분포 $\mathrm{Exp}(\lambda^2/2)$를 따르는 보조변수에 대한 정규분포의 스케일혼합(scale mixture) 형태를 사용하였다.

$$Laplace(\widetilde{\theta}_j | 0, \frac{\lambda}{\sigma}) = \int_0^\infty N(\widetilde{\theta}_j | 0, \sigma^2\tau_j^2)\, \mathrm{Exp}(\tau_j^2 \,|\, \frac{\lambda^2}{2})\, d\tau \tag{3}$$

1.3. 해석적 형식(analytic form)으로 구하기 어려운 사후확률

각 설명변수가 회귀모형에 포함될 사후확률을 구할 수 있어야 하나 해석적 형식으로 도출

하기 어렵기 때문에 MCMC알고리즘을 활용하여 생성한 샘플들로 비편향 추정을 할 수 있다. 그러나 MCMC는 데이터가 방대하고 설명변수가 많아질수록 마르코프 연쇄(markov chain)가 평형(equilibrium)에 이르는데 필요한 시간이 기하급수적으로 늘어나며 수렴성도 판단하기 어렵다. 고차원의 모형과 방대한 데이터에 강건한 다른 베이지안 추론법을 다음 장에서 소개한다.

2. 변분추론 (Variational Inference, VI)

변분추론은 목표분포의 수식을 해석적 형식으로 구하기 어려울 때 목표분포와 근접한 변분분포(variational distribution)를 구하는 최적화(optimization) 기법이다. 관측변수들 $\boldsymbol{o}=(o_1,...,o_R)$과 잠재변수들 $\boldsymbol{z}=(z_1,...,z_K)$이 주어져 있다고 가정하자. VI는 목적함수(objective function)를 사후분포 $\pi(\boldsymbol{z}|\boldsymbol{o})$와 변분분포 $q(\boldsymbol{z})$간 비유사성 정도를 측정하는 Kullack-Leibler (KL) 발산(divergence)으로 다음과 같이 정의하고 이를 최소화하는 최적의 변분분포를 구한다.

$$q^{\star}(\boldsymbol{z}) = \arg\min_{q\in Q} \mathrm{KL}(q(\boldsymbol{z})\|\pi(\boldsymbol{z}|\boldsymbol{o})) = \arg\min_{q\in Q} \mathrm{E}_{q(\boldsymbol{z})}\left[\log\frac{q(\boldsymbol{z})}{\pi(\boldsymbol{z}|\boldsymbol{o})}\right] \tag{4}$$

그러나 사후분포 $\pi(\boldsymbol{z}|\boldsymbol{o})$을 알 수 없기에 직접적으로 KL 발산을 계산하는 것은 불가능하다. 따라서 대체 목적함수인 ELBO를 최대화하는 문제로 치환한다.

$$\begin{aligned} q^{\star}(\boldsymbol{z}) &= \arg\min_{q\in Q} \mathrm{E}_{q(\boldsymbol{z})}\left[\log\frac{q(\boldsymbol{z})}{\pi(\boldsymbol{z}|\boldsymbol{o})}\right] \\ &= \arg\min_{q\in Q} \mathrm{E}_{q(\boldsymbol{z})}\left[\log\frac{q(\boldsymbol{z})}{\pi(\boldsymbol{z},\boldsymbol{o})}\right] - \log\pi(\boldsymbol{o}) \\ &= \arg\max_{q\in Q} \underbrace{\mathrm{E}_{q(\boldsymbol{z})}\left[\log\frac{\pi(\boldsymbol{z},\boldsymbol{o})}{q(\boldsymbol{z})}\right]}_{\mathrm{ELBO}} \end{aligned} \tag{5}$$

2.1. 평균장 변분추론(Mean Field Variational Inference)

최적의 변분분포는 사후분포와 달리 계산이 가능(tractable)해야하기 때문에 변분분포 후보군에 제약을 가할 수 있다. 특히 평균장 변분추론의 경우 제한된 분포족(distribution family)은 모든 잠재변수들이 완전한 독립(fully independence)이라는 가정을 가지며

$$Q = \left\{q : q(z_1,\cdots,z_K) = \prod_{k=1}^{K} q(z_k)\right\} \tag{6}$$

변분 미적분(variational calculus)을 통해 최적 변분분포를 다음과 같은 꼴로 구할 수 있다.

$$q^{\star}(z_k) \propto \exp \mathrm{E}_{q(\boldsymbol{z}_{-k})}[\log p(z_k|\boldsymbol{z}_{-k},\boldsymbol{o})] \tag{7}$$

$\boldsymbol{z}_{-k}$은 잠재변수들 $\boldsymbol{z}$에서 k번째 변수를 제외한 변수들이다. 평균장 변분추론의 가정에 따라 앞서 설명한 모형의 잠재변수들 $\boldsymbol{z}=(\rho,\alpha,\tilde{\boldsymbol{\theta}},\boldsymbol{\tau^2},\sigma^2,\boldsymbol{\gamma})$의 변분분포를 다음과 같이 분해(factorize)할 수 있다.

$$q(\boldsymbol{z}) = q(\alpha)q(\tilde{\boldsymbol{\theta}})q(\boldsymbol{\tau^2})q(\rho)q(\sigma^2)\prod_{j=1}^{p} q(\gamma_j) \tag{8}$$

각 잠재변수에 대한 변분분포의 식과 유도 과정, 그리고 Coordinate Ascent Algorithm (CAA)을 활용한 변분분포 업데이트 과정은 부록에 서술하였다.

2.2. 구조화 변분추론(Structured Variational Inference, SVI)

평균장 변분추론은 MCMC 알고리즘에 비해 빠르고 효율적인 계산이 가능하지만, 완전한 독립 가정으로 인해 편향이 존재하고 목적함수 ELBO의 해집합인 변분분포족을 제한함으로써 목적함수가 볼록하지 않으며(nonconvexity) 다수의 지역 최적값(local optima)을 가진다는 단점이 있다. 평균장 변분추론의 완전한 독립 가정을 완화시킨 것이 SVI이며 사후분포에 더 정확하게 근사하게 할 수 있을뿐더러 지역 최적화 문제에 더 강건하다고 알려져 있다. 그러나 SVI는 구조화된 변분분포(structured variational distribution)가 계산이 가능해야 한다는 점에서 방법론이 제한된다(Hoffman and Blei, 2015).

이전에 제시된 변분추론을 활용한 베이지안 변수선택법들은 공통적으로 $\boldsymbol{\gamma}$의 변분분포가 완전히 분해된다는 평균장 가정을 하였다(Carbonetto and Stephens, 2012, Huang et al., 2016, Ormerod et al., 2017, and Ray and Szabó, 2022). 완전한 독립성 가정으로 인해 각 변분분포 $q(\gamma_j)$의 식을 해석적으로 도출할 수 있지만 설명변수의 개수가 많아질수록 실제 사후분포와의 불일치가 심화될 수 있다. 따라서 이 논문에서는 이진 변수들간의 통계적 의존성도 함께 고려한 구조화된 변분분포 $q(\boldsymbol{\gamma})$를 활용할 것이다.

2.3. 구조화된 변분분포의 계산이 어려운 기댓값

$\boldsymbol{\gamma}$의 구조화된 변분분포는 다음과 같이 이차 형식(quadratic form)으로 유도할 수 있으며

$$q(\boldsymbol{\gamma}) = \frac{Q(\boldsymbol{\gamma})}{Z} = \frac{\exp(\boldsymbol{\psi}^\top \boldsymbol{\gamma} + \boldsymbol{\gamma}^\top \boldsymbol{\Psi} \boldsymbol{\gamma})}{Z} \tag{9}$$

$$Z = \sum_{\boldsymbol{\gamma}} Q(\boldsymbol{\gamma}) \tag{10}$$

자세한 유도과정은 부록에 서술하였다. VI를 이행하기 위해 기댓값 $\mathrm{E}_{q(\boldsymbol{\gamma})}[\cdot]$을 구할 수 있어야 하나 2^p만큼의 셀 수 없이 많은 조합의 합계를 계산해야 한다. 이를 해결하기 위해 깁스 샘플링(Gibbs Sampling)을 활용하여 기댓값을 근사할 수 있지만(Mimno et al., 2012), 깁스 샘플링은 변수들 간에 큰 상관관계를 가지는 경우나 다봉분포처럼 복잡한 분포에 대해 local mode에 갇히는 경향이 있어 정확한 기댓값 추정이 어려울 수 있다. 이러한 한계를 극복하기 위해 Sequential Monte Carlo Sampler(SMCS, Del Moral et al., 2006)를 VI에 중첩된 형태로 활용할 것이며 이후 3장에서 소개한다.

3. SMCS 알고리즘

SMCS는 목표분포(target distribution)의 기댓값을 근사하기 위해 샘플링을 기반으로 한 Monte Carlo 기법이되 중요도 샘플링(Importance Sampling, IS)과 중간분포들(intermediate distributions), MCMC 전이(transition), 그리고 resampling을 활용하여 근사의 정확도를 높이는 방법이다. 기댓값을 근사하기 위해 IS나 MCMC를 단독으로 사용할 수 있지만 SMCS는 각각이 가지는 단점을 보완하면서 추가로 중간 분포들과 resampling을 활용하기 때문에 근사의 정확성을 높일 수 있다는 장점을 가진다.

3.1. 중요도 샘플링

어떤 함수 $g(\cdot)$에 대한 목표 분포 $q(\boldsymbol{\gamma})$의 기댓값 $\mathrm{E}_{q(\boldsymbol{\gamma})}[g(\boldsymbol{\gamma})]$을 구하고자 하되 목표분포의 정규화 상수를 구하기 어렵고 비정규분포(unnormalized distribution) $Q(\boldsymbol{\gamma})$의 식을 알고 있다고 가정하자. IS는 샘플링하기 쉬운 제안분포(proposal distribution) $\eta(\boldsymbol{\gamma})$로부터 S개의

샘플들을 생성하고

$$\boldsymbol{\gamma}^{(1)},\ldots,\boldsymbol{\gamma}^{(S)} \overset{iid}{\sim} \eta(\boldsymbol{\gamma}) \tag{11}$$

각 샘플에 대한 가중치를 활용하여 기댓값 $\mathrm{E}_{q(\boldsymbol{\gamma})}[g(\boldsymbol{\gamma})]$을 추정한다.

$$\begin{aligned} \mathrm{E}_{q(\boldsymbol{\gamma})}[g(\boldsymbol{\gamma})] &\approx \sum_{s=1}^{S} \widetilde{W}^{(s)} g(\boldsymbol{\gamma}^{(s)}), \\ \text{where} \quad \widetilde{W}^{(s)} &= \frac{W^{(s)}}{\sum_{s'} W^{(s')}} \quad \text{and} \quad W^{(s)} = \frac{Q(\boldsymbol{\gamma}^{(s)})}{\eta(\boldsymbol{\gamma}^{(s)})} \end{aligned} \tag{12}$$

추정값의 정확도를 높이기 위해서 목표분포에 가까운 제안분포를 사용해야 하나 분포의 차원이 커질수록 목표분포와 유사한 제안분포 찾는 것이 어려워진다.

3.2. 중간분포와 MCMC 전이

SMCS는 목표분포로써 $q(\boldsymbol{\gamma})$를 직접적으로 사용하기보다 $q_0(\boldsymbol{\gamma}) = \eta(\boldsymbol{\gamma})$와 $q_T(\boldsymbol{\gamma}) = q(\boldsymbol{\gamma})$ 사이를 연결하는 $T-1$개의 중간분포들 $\{q_t(\boldsymbol{\gamma})\}_{t=1}^{T-1}$을 활용하여 결합분포수열 $\{\tilde{q}_t(\boldsymbol{\gamma}_{0:t})\}_{t=0}^{T}$을 만들고 각 결합분포 $\tilde{q}_t(\boldsymbol{\gamma}_{0:t})$을 매 t번째 시점마다의 목표분포로 사용한다. $\tilde{q}_t(\boldsymbol{\gamma}_{0:t})$는 주변분포로써 $q_t(\boldsymbol{\gamma}_t) = Q_t(\boldsymbol{\gamma})/Z_t$을 가지며 다음과 같이 정의된다.

$$\begin{aligned} \tilde{q}_t(\boldsymbol{\gamma}_{0:t}) &= \frac{\widetilde{Q}_t(\boldsymbol{\gamma}_{0:t})}{Z_t}, \\ \text{where} \quad \widetilde{Q}_t(\boldsymbol{\gamma}_{0:t}) &= Q_t(\boldsymbol{\gamma}_t) \prod_{l=0}^{t-1} L_l(\boldsymbol{\gamma}_l | \boldsymbol{\gamma}_{l+1}) \end{aligned} \tag{13}$$

$L_l(\boldsymbol{\gamma}_l|\boldsymbol{\gamma}_{l+1})$은 $\boldsymbol{\gamma}_{l+1}$에서 $\boldsymbol{\gamma}_l$로 이동시키는 인공 Backward Markov Kernel(BMK)이다. 제안분포로써 $\boldsymbol{\gamma}_{l-1}$에서 $\boldsymbol{\gamma}_l$로 이동시키는 Forward Markov Kernel(FMK) $K_l(\boldsymbol{\gamma}_l|\boldsymbol{\gamma}_{l-1})$들로 정의되는 결합분포

$$\eta_t(\boldsymbol{\gamma}_{1:t}) = q_0(\boldsymbol{\gamma}_0) \prod_{l=1}^{t} K_l(\boldsymbol{\gamma}_l | \boldsymbol{\gamma}_{l-1}) \tag{14}$$

를 활용하여 각 샘플에 대한 IS 비정규 가중치는 다음과 같이 계산한다.

$$W_t^{(s)} \propto \frac{\tilde{q}_t(\boldsymbol{\gamma}_{0:t}^{(s)})}{\eta_t(\boldsymbol{\gamma}_{0:t}^{(s)})} = \frac{q_t(\boldsymbol{\gamma}_t^{(s)}) \prod_{l=0}^{t-1} L_l(\boldsymbol{\gamma}_l^{(s)} | \boldsymbol{\gamma}_{l+1}^{(s)})}{\eta_0(\boldsymbol{\gamma}_0^{(s)}) \prod_{l'=1}^{t} K_{l'}(\boldsymbol{\gamma}_{l'}^{(s)} | \boldsymbol{\gamma}_{l'-1}^{(s)})} \propto W_{t-1}^{(s)} w_t(\boldsymbol{\gamma}_{t-1}^{(s)}, \boldsymbol{\gamma}_t^{(s)}) \tag{15}$$

$$w_t(\boldsymbol{\gamma}_{t-1}^{(s)}, \boldsymbol{\gamma}_t^{(s)}) = \frac{Q_t(\boldsymbol{\gamma}_t^{(s)}) L_{t-1}(\boldsymbol{\gamma}_{t-1}^{(s)} | \boldsymbol{\gamma}_t^{(s)})}{Q_{t-1}(\boldsymbol{\gamma}_{t-1}^{(s)}) K_t(\boldsymbol{\gamma}_t^{(s)} | \boldsymbol{\gamma}_{t-1}^{(s)})} \tag{16}$$

$q_t(\boldsymbol{\gamma}_t)$는 $\tilde{q}_t(\boldsymbol{\gamma}_{0:t})$의 주변분포이기 때문에 정규화된 $\{\widetilde{W}_t^{(s)}\}_{s=1}^{S}$를 활용하여 분포를 근사 가능하며

$$q_t^N(d\boldsymbol{\gamma}) = \sum_{s=1}^{S} \widetilde{W}_t^{(s)} \delta_{\boldsymbol{\gamma}_t^{(s)}}(d\boldsymbol{\gamma}) \tag{17}$$

함수 g에 대한 기댓값도 근사 가능하다.

$$\mathrm{E}_{q_t^N}[g(\boldsymbol{\gamma})] = \sum_{s=1}^{(s)} \widetilde{W}_t^{(s)} g(\boldsymbol{\gamma}_t^{(s)}) \tag{18}$$

가중치들의 분산을 최소화하기 위한 BMK 설계가 중요한데, FMK을 활용하여 설계한다면

최적의 BMK는 다음과 같이 계산될 수 있지만

$$L_{t-1}^{opt}(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\gamma}_t) = \frac{\eta_{t-1}(\boldsymbol{\gamma}_{t-1})K_t(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{t-1})}{\eta_t(\boldsymbol{\gamma}_t)} \tag{19}$$

그러나 주변분포를 구하기가 어렵다. 이때 FMK로써 $q_t(\boldsymbol{\gamma}_t)$에 불변한 MCMC kernel을 활용하여 식 (17)을 근사한다면 다음과 같은 식을 고려할 수 있으며

$$L_{t-1}(\boldsymbol{\gamma}_{t-1}|\boldsymbol{\gamma}_t) = \frac{q_t(\boldsymbol{\gamma}_{t-1})K_t(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{t-1})}{q_t(\boldsymbol{\gamma}_t)} \tag{20}$$

만약 $q_t(\boldsymbol{\gamma}_t)$와 $q_{t-1}(\boldsymbol{\gamma}_{t-1})$의 차이가 작다면 식 (17)에 대한 좋은 근사가 될 수 있다. 중간분포로 목표분포와 제안분포 사이에 어닐링을 준 분포로 결정한다면

$$\begin{gathered} Q_t(\boldsymbol{\gamma}) \propto Q_0(\boldsymbol{\gamma})^{1-\beta_t} Q_T(\boldsymbol{\gamma})^{\beta_t}, \\ \text{where } \ 0 = \beta_0 < \cdots < \beta_T = 1 \end{gathered} \tag{21}$$

식 (16)은 다음과 같이 표현될 수 있다.

$$w_t(\boldsymbol{\gamma}_{t-1}^{(s)}, \boldsymbol{\gamma}_t^{(s)}) = \frac{Q_t(\boldsymbol{\gamma}_{t-1}^{(s)})}{Q_{t-1}(\boldsymbol{\gamma}_{t-1}^{(s)})} = \left(\frac{Q_T(\boldsymbol{\gamma}_{t-1}^{(s)})}{Q_0(\boldsymbol{\gamma}_{t-1}^{(s)})} \right)^{\beta_t - \beta_{t-1}} \tag{22}$$

SMCS의 알고리즘 순서는 아래와 같다.

- 각 샘플 $\boldsymbol{\gamma}_0^{(s)}$을 $q_0(\boldsymbol{\gamma})$에서 생성한다.
- FOR $t = 0 : T-1$,
  - 샘플들 $\{\boldsymbol{\gamma}_t^{(s)}\}_{s=1}^S$의 가중치 $\{W_t^{(s)}\}_{s=1}^S$를 계산하고 정규화시킨다.
  - 만약 ESS_t값이 특정 threshold보다 작다면 정규화된 가중치 $\{\widetilde{W}_t^{(s)}\}_{s=1}^S$을 바탕으로 resampling을 한다.
  - 새로운 샘플들 $\{\boldsymbol{\gamma}_{t+1}^{(s)}\}_{s=1}^S$을 중간분포 $q_t(\boldsymbol{\gamma}) \propto Q_t(\boldsymbol{\gamma})$에 불변한 FMK K_t 사용하여 샘플링한다, i.e., $\boldsymbol{\gamma}_{t+1}^{(s)} \sim K_{t+1}(\cdot|\boldsymbol{\gamma}_t^{(s)})$

3.3. ESS_t와 resampling의 역할

이때 각 t시점마다 resampling을 하기 위한 조건을 구하기 위한 effect sample size (ESS)는 다음과 같이 계산한다.

$$\mathrm{ESS}_t = \frac{\left(\sum_{s=1}^S W_t^{(s)} \right)^2}{\sum_{s=1}^S \left(W_t^{(s)} \right)^2} \tag{23}$$

ESS_t는 가중치들 $\{W_t^{(s)}\}_{s=1}^S$의 분산을 측정하는데 사용되며 ESS_t값이 작을수록 $\{\boldsymbol{\gamma}_t^{(s)}\}_{s=1}^S$의 분산이 커서 기댓값을 근사하기 위한 샘플들 $\{\boldsymbol{\gamma}_t^{(s)}\}_{s=1}^S$에서 degeneracy가 일어날 여지가 있다. 따라서 ESS_t값이 미리 명시한 값보다 작은 경우 낮은 가중치의 샘플들을 제거하고 높은 가중치의 샘플들만을 복제하여 활용하기 위해 resampling을 수행한다.

3.4. 효율성을 높이기 위한 과거 샘플들의 재활용

SMCS를 활용하여 $\mathrm{E}_{q(\boldsymbol{\gamma})}[g(\boldsymbol{\gamma})]$을 근사한다면 T시점의 샘플들의 가중치들 $\{W_t^{(s)}\}_{s=1}^S$만이 사용된다. 그러나 Gramacy et al. (2010)와 유사한 방법으로 T시점 이전의 샘플들도 활용하

여 기댓값을 근사할 수 있다. t시점의 샘플들 $\{\boldsymbol{\gamma}_t^{(s)}\}_{s=1}^{S}$에 대하여 새로운 비정규 가중치 $W_t^{\star}$를 정의하여 $\mathrm{E}_{q(\boldsymbol{\gamma})}[g(\boldsymbol{\gamma})]$에 대한 추정량 $\hat{g}_t$을 구할 수 있다.

$$\begin{aligned}\hat{g}_t &= \sum_{s=1}^{S} \frac{W_t^{\star\,(s)}}{\sum_{s'=1}^{S} W_t^{\star\,(s')}} g(\boldsymbol{\gamma}_t^{(s)}), \\ \text{where}\quad W_t^{\star\,(s)} &= W_t^{(s)} \frac{Q_T(\boldsymbol{\gamma}_t^{(s)})}{Q_t(\boldsymbol{\gamma}_t^{(s)})}\end{aligned} \tag{24}$$

총 T개의 추정량 $\hat{g}_0,\ldots,\hat{g}_{T-1}$들의 convex 합으로 $\mathrm{E}_{q(\boldsymbol{\gamma})}[g(\boldsymbol{\gamma})]$에 대한 추정량 $\hat{g}$을 정의할 수 있으며

$$\begin{aligned}\hat{g} &= \sum_{t=0}^{T-1} \lambda_t \hat{g}_t = \sum_{t=0}^{T-1}\sum_{s=1}^{S} \left(\lambda_t \frac{W_t^{\star\,(s)}}{\sum_{s'=1}^{S} W_t^{\star\,(s')}} \right) g(\boldsymbol{\gamma}_t^{(s)}) \\ \text{where}\quad 0 &\le \lambda_t \le \sum_{t=0}^{T-1} \lambda_t = 1\end{aligned} \tag{25}$$

effective sample size $\mathrm{ESS}^{\star}$는 다음과 같이 계산할 수 있다.

$$\mathrm{ESS}^{\star} = \left[\sum_{t=0}^{T-1}\sum_{s=1}^{S} \left(\lambda_t \frac{W_t^{\star\,(s)}}{\sum_{s'=1}^{S} W_t^{\star\,(s')}} \right)^2 \right]^{-1} \tag{26}$$

이를 최대화하는 각 최적의 λ_t은 라그랑주 승수법(lagrange multiplier)을 활용하여 구할 수 있다.

$$\lambda_t = \frac{\nu_t}{\sum_{t'=0}^{T-1} \nu_{t'}} \quad \text{where} \quad \nu_t = \frac{\left(\sum_{s=1}^{S} W_t^{\star\,(s)} \right)^2}{\sum_{s'=1}^{S} \left(W_t^{\star\,(s')} \right)^2} \tag{27}$$

4. SVI 안에 중첩된 SMCS (SVI-S)

SVI-S의 특징은 두 계층의 중간분포들을 활용하여 구조화 변분분포 $q(\boldsymbol{\gamma})$의 수렴분포를 효과적으로 구하는 것이다. 변분추론의 CAA에서 $\boldsymbol{\gamma}$의 초기변분분포를 $q_{[0]}(\boldsymbol{\gamma})$라고 하고 총 I번의 iteration 이후 수렴한 변분분포를 $q_{[I]}(\boldsymbol{\gamma})$라고 하자. 첫 번째 계층의 중간분포들은 변분추론 알고리즘이 수렴하기 이전에 생성된 분포들이며, i.e., $\{q_{[i]}(\boldsymbol{\gamma})\}_{i=1}^{I-1}$ 두 번째 계층의 중간분포들은 각 $i+1$번째 비정규 변분분포 $Q_{[i+1]}(\boldsymbol{\gamma})$와 i번째 비정규 변분분포 $Q_{[i]}(\boldsymbol{\gamma})$ 사이를 연결하는 어닐링이 된 형태에 비례한다.

$$\begin{aligned}Q_{[i],t}(\boldsymbol{\gamma}) &\propto (Q_{[i]}(\boldsymbol{\gamma}))^{1-\beta_t} (Q_{[i+1]}(\boldsymbol{\gamma}))^{\beta_t}, \\ &= \exp\left\{ \left((1-\beta_t)\psi_{[i]} + \beta_t \psi_{[i+1]} \right)^{\top} \boldsymbol{\gamma} + \boldsymbol{\gamma}^{\top} \left((1-\beta_t)\Psi_{[i]} + \beta_t \Psi_{[i+1]} \right) \boldsymbol{\gamma} \right\} \\ &\text{where}\quad 0 = \beta_0 < \cdots < \beta_T = 1\end{aligned} \tag{28}$$

이에 따라 i번째 변분분포의 t번째 어닐링이 된 각 가중치 $W_{[i],t}^{(s)}$은 다음과 같이 연쇄적으로 계산할 수 있다.

$$W_{[0],t}^{(s)} = \prod_{l=1}^{t} \frac{Q_{[0],l}(\boldsymbol{\gamma}_{[0],l}^{(s)})}{Q_{[0],l-1}(\boldsymbol{\gamma}_{[0],l}^{(s)})},$$

$$W_{[1],t}^{(s)} = W_{[0],T}^{(s)} \prod_{l=1}^{t} \frac{Q_{[1],l}(\boldsymbol{\gamma}_{[1],l}^{(s)})}{Q_{[1],l-1}(\boldsymbol{\gamma}_{[1],l}^{(s)})}, \tag{29}$$

$$W_{[i],t}^{(s)} = W_{[i-1],T}^{(s)} \prod_{l=1}^{t} \frac{Q_{[i],l}(\boldsymbol{\gamma}_{[i],l}^{(s)})}{Q_{[i],l-1}(\boldsymbol{\gamma}_{[i],l}^{(s)})}$$

$\{W_{[i],T}^{(s)}\}_{s=1}^{S}$은 $q_{[i+1]}(\boldsymbol{\gamma})$의 기댓값을 계산하기 위해 사용되는 최종 가중치 집합이며 이에 사용되는 샘플들 $\{\boldsymbol{\gamma}_{[i],t}^{(s)}\}_{s=1}^{S}$의 전이과정은 제안분포를 초기변분분포 $q_{[0]}(\boldsymbol{\gamma})$로 결정했을 때의 SMCS 전이과정과 유사하다.

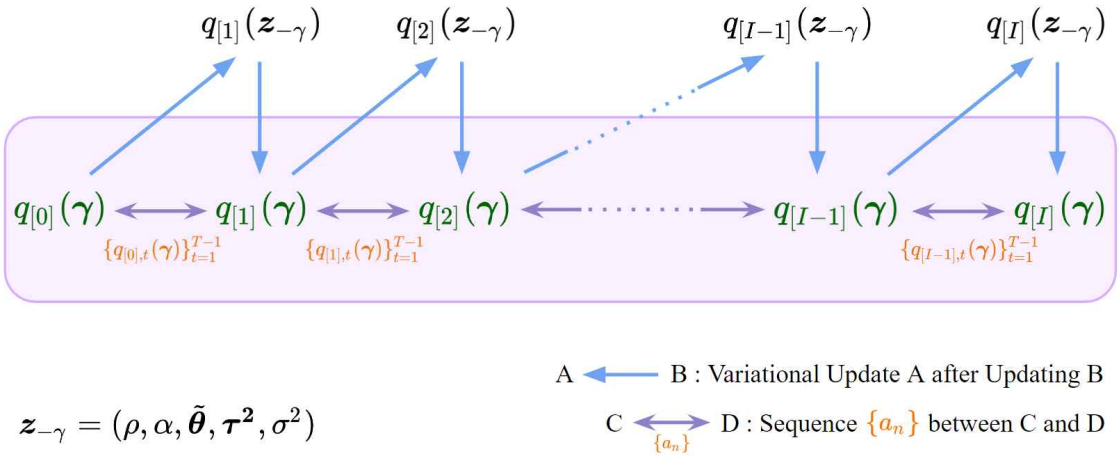

그림1. 두 계층의 중간분포 집합을 활용하는 SVI-S 알고리즘

4.1. 첫 번째 계층의 중간분포들에 대한 어닐링

SMCS는 제안분포와 목표분포의 차이가 클 때 중간분포 개수를 늘려서 기댓값의 근사 정확도를 높일 수 있지만 계산하는 시간이 더 소요되는 단점이 있다. 첫 번째 계층의 각 두 중간분포 $q_{[i]}(\boldsymbol{\gamma})$와 $q_{[i+1]}(\boldsymbol{\gamma})$의 차이도 클 수 있기 때문에 두 분포의 거리를 줄이기 위한 어닐링 파라미터 수열 $\{\xi_i\}_{i=1}^{I}$을 추가하여 첫 번째 계층의 중간분포를 다음과 같이 재정의하였다.

$$\begin{gathered} q_{[i]}(\boldsymbol{\gamma};\xi_i) \propto q_{[i]}(\boldsymbol{\gamma})^{\xi_i}, \\ \text{where } 0.001 = \xi_1 < \xi_2 < \cdots < \xi_u = 1 \\ \text{and } \xi_i = 1 \text{ for } u+1 \le i \le I \end{gathered} \tag{30}$$

이때 CAA가 수렴하기 전까지 생성된 수열 $\{q_{[i]}(\boldsymbol{\gamma};\xi_i)\}_{i=1}^{I}$은 ELBO에서 유도된 것이 아니기에 식 (19)에 대한 정당성이 필요하다. 사실 아래와 같은 새로운 목적함수에서 도출된 식임을 보일 수 있으며

$$\begin{aligned} q^{\star}(\boldsymbol{\gamma};\xi) &= \arg\max_{q \in Q} \mathrm{ELBO}(\xi) \\ &= \arg\max_{q \in Q} \left\{ \mathrm{E}_{q(\boldsymbol{z})}[\log \pi(\boldsymbol{z},\boldsymbol{y})] + \mathrm{E}_{q(\boldsymbol{z}_{-\gamma})}[-\log q(\boldsymbol{z}_{-\gamma})] + \frac{1}{\xi}\mathrm{E}_{q(\boldsymbol{\gamma})}[-\log q(\boldsymbol{\gamma})] \right\} \end{aligned} \tag{31}$$

이는 결정론 어닐링 변분추론(Deterministic Annealing Variational Inference)의 목적함수에 해당한다. 결정론 어닐링 변분추론에 대한 자세한 내용은 Katahira et al. (2007)을 참고

하기를 바란다.

SVI-S의 알고리즘은 그림1에 요약해놓았으며 자세한 과정은 부록에 서술하였다.

5. 실험

SVI-S의 변수선택 성능을 측정하기 위해 다음과 같은 인공데이터(synthetic data)를 생성하고 다른 6가지의 최적화기법에 기반한 방법론들과 비교하였다. n개의 관측치에 대하여 총 p개의 설명변수 중 s개의 설명변수가 회귀모형에 포함되고 나머지 $p-s$개의 설명변수는 회귀모형에 포함되지 않는다. 회귀모형에 포함되는 설명변수의 회귀계수들은 $[-10,-1]\cup[1,10]$에서 균등하게 생성되었다. 또한 모든 설명변수들은 서로 pairwise 상관계수 ϕ를 가진다.

변수선택의 성능을 측정하기 위한 평가기준으로 두 가지 지표를 사용하였는데, 한 가지는 실제로 모형에 포함되어야 할 설명변수를 포함되지 않을 변수로 오분류한 총 거짓음성(false negative, FN) 개수와 다른 한 가지는 실제로 모형에 포함되지 말아야 할 설명변수를 포함할 변수로 오분류한 총 거짓양성(false positive, FP) 개수이며 식은 아래와 같다.

$$\mathrm{FN}=\sum_{j=1}^{p}1(\mathrm{E}_{q(\boldsymbol{\gamma})}[\gamma_j|\boldsymbol{y}]<0.5,\gamma_j=1) \tag{32}$$

$$\mathrm{FP}=\sum_{j=1}^{p}1(\mathrm{E}_{q(\boldsymbol{\gamma})}[\gamma_j|\boldsymbol{y}]\geq 0.5,\gamma_j=0) \tag{33}$$

SVI-S의 셋팅으로 100개의 샘플들과 각 $q(\boldsymbol{\gamma})^{[i]}$에 대해 300개의 annealing 분포들, 그리고 MCMC transition을 위해 깁스 샘플링을 사용하였다. SVI-S와 비교하기 위해 사용한 다른 방법들로 EMVS (Ročková et al, 2014), SSLASSO (Ročková et al, 2018), varbvs (Carbonetto et al, 2012), sparsevb (Ray and Szabó, 2022), 평균장 변분추론 MFVI, 그리고 깁스 샘플링으로 근사한 구조화 변분추론 SVI-G이 있다.

네 가지의 (n,p,s,ϕ)조합에 대해 각각 100개의 랜덤 인공데이터를 생성하여 실험하였고 상자그림(boxplot)으로 그린 결과가 그림2부터 그림5까지이다. 다른 방법들과 비교했을 때 SVIS-S가 우수한 성능을 보이고 있음을 알 수 있다.

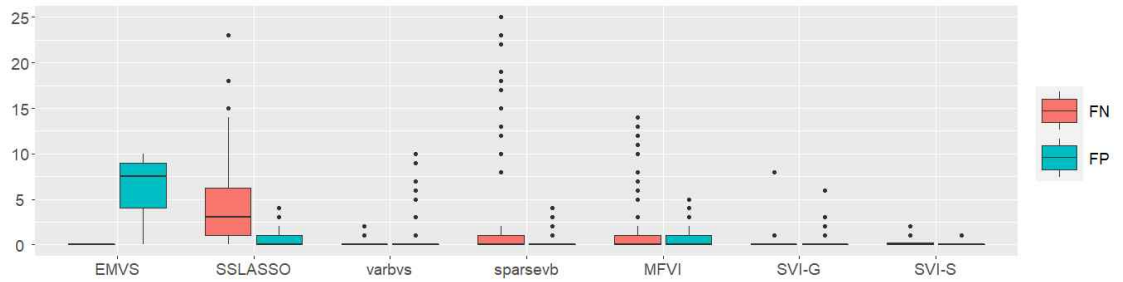

그림2. $(n,p,s,\phi)=(50,100,10,0.3)$에서의 FN과 FP의 상자그림

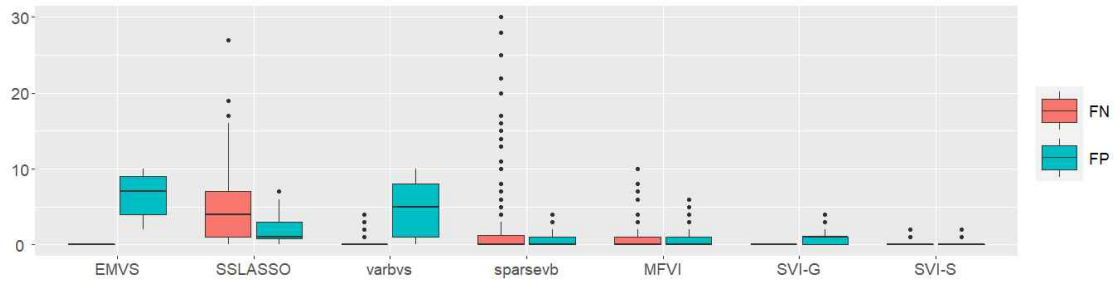

그림3. $(n,p,s,\phi)=(50,100,10,0.6)$에서의 FN과 FP의 상자그림

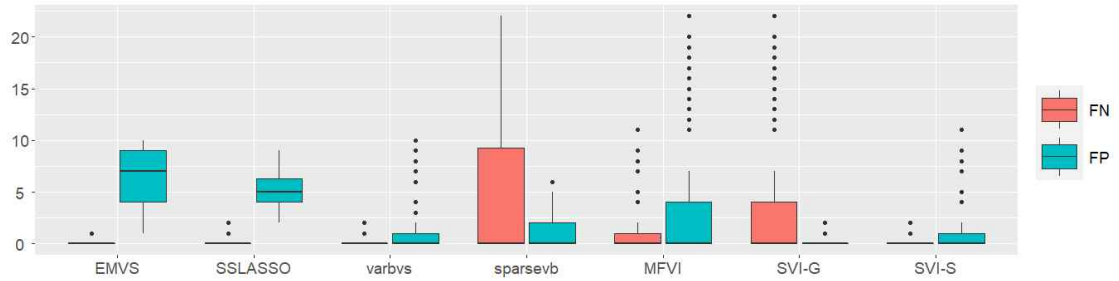

그림4. $(n,p,s,\phi)=(50,200,10,0.3)$에서의 FN과 FP의 상자그림

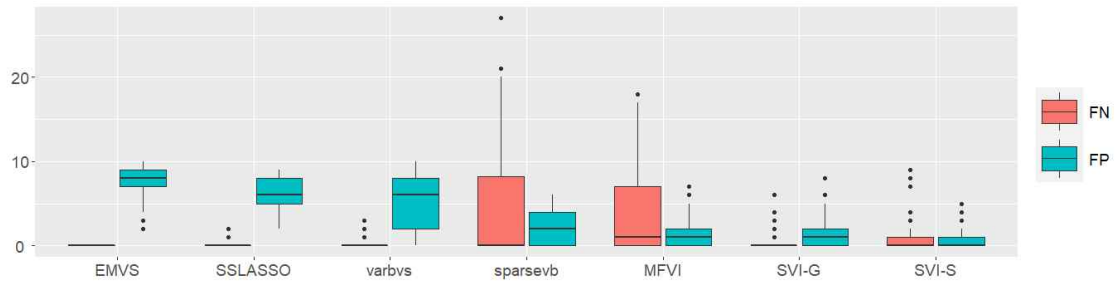

그림5. $(n,p,s,\phi)=(50,200,10,0.6)$에서의 FN과 FP의 상자그림

6. 결론 및 향후 연구

본 논문에서는 베이지안 변수선택을 위한 평균장 변분추론의 한계를 극복하기 위해 구조화 변분추론을 활용하였으며 계산 불가능한 정규화 상수 문제를 효과적으로 해결하기 위해 SMCS를 응용한 SVI-S를 제안하였다. 향후 계획으로 구조화된 변분분포를 보다 빠르고 정확하게 근사하기 위한 방법론을 개발할 것이며 구조화 변분추론의 수렴속도 및 최적화된 변분분포의 불확실성 등 이론적 성질을 유도하는 연구를 할 것이다.

7. 부록

n개의 관측치를 가진 반응변수벡터 $\boldsymbol{y}$와 p개의 설명변수벡터들을 가진 관측행렬(design matrix) $\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p]$에 대하여 선형회귀모형을 $\boldsymbol{\gamma}$와 $\tilde{\boldsymbol{\theta}}$에 대한 식으로 표현할 수 있다.

$$\pi(\boldsymbol{y}\,|\,\alpha,\boldsymbol{\gamma},\tilde{\boldsymbol{\theta}},\sigma^2)=N(\alpha\mathbf{1}_n+\boldsymbol{X\Gamma}\tilde{\boldsymbol{\theta}},\sigma^2\mathrm{I}_n)$$

$\mathbf{1}_n$은 모든 요소가 1로만 이루어진 n차원 벡터, $\boldsymbol{\Gamma}=diag(\gamma_1,\ldots,\gamma_p)$은 이진변수들의 대각행렬에 해당한다. Slab 변수들 $\tilde{\boldsymbol{\theta}}$의 사전분포는 다음과 같이 계층적으로 표현할 수 있다.

$$\pi(\tilde{\boldsymbol{\theta}}|\sigma^2,\boldsymbol{\tau^2})=N(\mathbf{0}_p,\sigma^2\boldsymbol{V})$$

$$\pi(\boldsymbol{\tau^2}|\lambda) = \prod_{j=1}^{p} \mathrm{Exp}\left(\frac{\lambda^2}{2}\right)$$

$\mathbf{0}_p$은 모든 요소가 0으로만 이루어진 p차원 벡터, $\boldsymbol{\tau^2} = (\tau_1^2, ..., \tau_p^2)^\top$은 보조변수 벡터, 그리고 $\boldsymbol{V} = diag(\boldsymbol{\tau^2})$은 보조변수들의 대각행렬에 해당한다.

7.1. 평균장 변분추론 가정하에서의 변분분포 유도

표기의 편의를 위해 잠재변수집합 $\boldsymbol{z} = (\rho, \alpha, \tilde{\boldsymbol{\theta}}, \boldsymbol{\tau}, \sigma^2, \boldsymbol{\gamma})$의 변분분포 $q(\boldsymbol{z})$에 대하여 어떤 확률변수 x를 제외하고 기댓값을 취한 값을 $\mathrm{E}_{-x}[\cdot]$라고 하자. 또한 기댓값을 취하는 분포대상이 자명한 경우 $\mathrm{E}_q[\cdot]$로 표기한다.

1) ρ의 변분분포 유도과정은 다음과 같다.

$$\log q^\star(\rho) \propto \sum_{j=1}^{p} \mathrm{E}_q[\gamma_j] \log\rho + (2p - \sum_{j=1}^{p} \mathrm{E}_q[\gamma_j] - 1)\log(1-\rho)$$

이에 따라 ρ는 아래의 베타분포를 따른다.

$$\begin{aligned} q^\star(\rho) &= Beta(A_\rho, B_\rho), \\ A_\rho &= 1 + \sum_{j=1}^{p} \mathrm{E}_q[\rho_j], \\ B_\rho &= 2p - \sum_{j=1}^{p} \mathrm{E}_q[\rho_j] \end{aligned}$$

다른 변분분포 계산에 필요한 기댓값 계산은 다음과 같다.

$$\begin{aligned} \mathrm{E}_q[\mathrm{logit}(\rho)] &= \mathrm{E}_q[\log(\rho)] - \mathrm{E}_q[\log(1-\rho)] \\ &= (D(A_\rho) - D(A_\rho + B_\rho)) - (D(B_\rho) - D(A_\rho + B_\rho)) \\ &= D(A_\rho) - D(B_\rho) \end{aligned}$$

$D(\cdot)$은 $digamma$함수이다.

2) α의 변분분포 유도과정은 다음과 같다.

$$\begin{aligned} \log q^\star(\alpha) &\propto \mathrm{E}_{-\alpha}\left[-\frac{1}{2\sigma^2} \| (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\Gamma}\tilde{\boldsymbol{\theta}}) - \alpha \mathbf{1}_n \|_2^2 \right] \\ &\propto -\frac{\mathrm{E}_q[\sigma^{-2}]}{2} \left(n\alpha^2 - 2(\boldsymbol{y} - \boldsymbol{X}\mathrm{E}_q[\boldsymbol{\Gamma}]\mathrm{E}_q[\tilde{\boldsymbol{\theta}}])^\top (\alpha \mathbf{1}_n) \right) \end{aligned}$$

이에 따라 α는 아래의 정규분포를 따른다.

$$\begin{aligned} q^\star(\alpha) &= N(\mu_\alpha, \sigma_\alpha^2), \\ \sigma_\alpha^2 &= (n\mathrm{E}_q[\sigma^{-2}])^{-1}, \\ \mu_\alpha &= \sigma_\alpha^2 \mathrm{E}_q[\sigma^{-2}] \mathbf{1}_n^\top (\boldsymbol{y} - \boldsymbol{X}\mathrm{E}_q[\boldsymbol{\Gamma}]\mathrm{E}_q[\tilde{\boldsymbol{\theta}}]) \end{aligned}$$

다른 변분분포 계산에 필요한 기댓값 계산은 다음과 같다.

$$\mathrm{E}_q[\alpha^2] = \sigma_\alpha^2 + \mu_\alpha^2$$

3) $\tilde{\boldsymbol{\theta}}$의 변분분포 유도과정은 다음과 같다.

$$\begin{aligned} \log q^\star(\tilde{\boldsymbol{\theta}}) &\propto \mathrm{E}_{-\tilde{\boldsymbol{\theta}}}\left[-\frac{1}{2\sigma^2} \left(\tilde{\boldsymbol{\theta}}^\top \boldsymbol{V}^{-1} \tilde{\boldsymbol{\theta}} - \| (\boldsymbol{y} - \alpha \mathbf{1}_n) - \boldsymbol{X}\boldsymbol{\Gamma}\tilde{\boldsymbol{\theta}} \|_2^2 \right) \right] \\ &\propto -\frac{\mathrm{E}_q[\sigma^2]}{2} \left(\tilde{\boldsymbol{\theta}}^\top (\mathrm{E}_q[\boldsymbol{V}^{-1}] + \boldsymbol{X}^\top \boldsymbol{X} \odot \mathrm{E}_q[\boldsymbol{\gamma}\boldsymbol{\gamma}^\top]) \tilde{\boldsymbol{\theta}} - 2\boldsymbol{y}^\top (\boldsymbol{X}\mathrm{E}_q[\boldsymbol{\Gamma}]\mathrm{E}_q[\tilde{\boldsymbol{\theta}}]) \right) \end{aligned}$$

이에 따라 $\tilde{\boldsymbol{\theta}}$는 아래의 정규분포를 따른다.

$$
\begin{aligned}
q^{\star}(\tilde{\boldsymbol{\theta}}) &= N(\boldsymbol{\mu}_{\tilde{\boldsymbol{\theta}}}, \Sigma_{\tilde{\boldsymbol{\theta}}}), \\
\Sigma_{\tilde{\boldsymbol{\theta}}} &= \mathrm{E}_q[\sigma^{-2}](\boldsymbol{X}^{\top}\boldsymbol{X} \odot \mathrm{E}_q[\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}}] + \mathrm{E}_q[\boldsymbol{V}^{-1}])^{-1}, \\
\boldsymbol{\mu}_{\tilde{\boldsymbol{\theta}}} &= \mathrm{E}_q[\sigma^{-2}]\Sigma_{\tilde{\boldsymbol{\theta}}}\mathrm{E}_q[\boldsymbol{\Gamma}]\boldsymbol{X}^{\top}\boldsymbol{y}
\end{aligned}
$$

다른 변분분포 계산에 필요한 기댓값 계산은 다음과 같다.

$$
\mathrm{E}_q[\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^{\mathsf{T}}] = \Sigma_{\tilde{\boldsymbol{\theta}}} + \boldsymbol{\mu}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{\mu}_{\tilde{\boldsymbol{\theta}}}^{\top}
$$

4) 각 τ_j^2의 변분분포 유도과정은 다음과 같다.

$$
\log q^{\star}(\tau_j) \propto -\frac{1}{2}\log\tau_j^2 - \frac{1}{2}\left(\lambda^2\tau_j^2 + \frac{\mathrm{E}_q[\sigma^{-2}]\mathrm{E}_q[\tilde{\theta}_j^2]}{\tau_j^2}\right)
$$

이에 따라 τ_j^2은 아래의 Generalized Inverse Gaussian (GIG)분포를 따른다.

$$
\begin{aligned}
q^{\star}(\tau_j^2) &= GIG(p_{\tau}, a_{\tau}, b_{\tau_j}), \\
p_{\tau} &= \frac{1}{2}, a_{\tau} = \lambda^2, b_{\tau_j} = \mathrm{E}_q[\sigma^{-2}]\mathrm{E}_q[\tilde{\theta}_j^2]
\end{aligned}
$$

다른 변분분포 계산에 필요한 기댓값 계산은 다음과 같다.

$$
\begin{aligned}
\mathrm{E}_q[\tau_j^{-2}] &= \frac{\sqrt{a_{\tau}}\boldsymbol{K}_{p_{\tau}+1}(\sqrt{a_{\tau}b_{\tau_j}})}{\sqrt{b_{\tau_j}}\boldsymbol{K}_{p_{\tau}}(\sqrt{a_{\tau}b_{\tau_j}})} - \frac{2p_{\tau}}{b_{\tau_j}}, \\
\mathrm{E}_q[\boldsymbol{V}^{-2}] &= diag(\mathrm{E}_q[\tau_1^2], \ldots, \mathrm{E}_q[\tau_p^2])
\end{aligned}
$$

$\boldsymbol{K}_p$는 modified Bessel function of the second kind에 해당한다.

5) σ^2의 변분분포 유도과정은 다음과 같다.

$$
\begin{aligned}
\log q^{\star}(\sigma^2) &\propto \mathrm{E}_{-\sigma^2}\left[-\frac{(n+p+1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\tilde{\boldsymbol{\theta}}^{\top}\boldsymbol{V}^{-1}\tilde{\boldsymbol{\theta}} + \|(\boldsymbol{y}-\alpha\mathbf{1}_n) - \boldsymbol{X}\boldsymbol{\Gamma}\tilde{\boldsymbol{\theta}}\|_2^2\right)\right] \\
&\propto -\frac{(n+p+1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\{\boldsymbol{y}^{\top}\boldsymbol{y} - 2\mathrm{E}_q[\alpha]\boldsymbol{y}^{\top}\mathbf{1} + n\mathrm{E}_q[\alpha^2] - 2(\boldsymbol{y}-\mathrm{E}_q[\alpha]\mathbf{1}_n)^{\top}\boldsymbol{X}\mathrm{E}_q[\boldsymbol{\Gamma}]\mathrm{E}_q[\tilde{\boldsymbol{\theta}}] \\
&\quad + tr\left((\boldsymbol{X}^{\top}\boldsymbol{X} \odot \mathrm{E}_q[\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}}] + \boldsymbol{V}^{-1})\mathrm{E}_q[\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^{\top}]\right)\}
\end{aligned}
$$

이에 따라 σ^2은 아래의 역감마분포를 따른다.

$$
\begin{aligned}
q^{\star}(\sigma^2) &= IG(A_{\sigma^2}, B_{\sigma^2}), \\
A_{\sigma^2} &= \frac{n+p}{2} \\
B_{\sigma^2} &= \{\boldsymbol{y}^{\top}\boldsymbol{y} - 2\mathrm{E}_q[\alpha]\boldsymbol{y}^{\top}\mathbf{1} + n\mathrm{E}_q[\alpha^2] - 2(\boldsymbol{y}-\mathrm{E}_q[\alpha]\mathbf{1}_n)^{\top}\boldsymbol{X}\mathrm{E}_q[\boldsymbol{\Gamma}]\mathrm{E}_q[\tilde{\boldsymbol{\theta}}] \\
&\quad + tr\left((\boldsymbol{X}^{\top}\boldsymbol{X} \odot \mathrm{E}_q[\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}}] + \boldsymbol{V}^{-1})\mathrm{E}_q[\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^{\top}]\right)\}/2
\end{aligned}
$$

다른 변분분포 계산에 필요한 기댓값 계산은 다음과 같다.

$$
\mathrm{E}_q[\sigma^{-2}] = \frac{A_{\sigma^2}}{B_{\sigma^2}}
$$

6) 각 γ_j의 변분분포 유도과정은 다음과 같다.

$$\begin{aligned}\log q^{\star}(\gamma_j) &\propto \mathrm{E}_{-\gamma_j}\left[\gamma_j \mathrm{logit}(\rho) - \frac{1}{2\sigma^2}\|(\boldsymbol{y}-\alpha\mathbf{1}_n) - \boldsymbol{X}\boldsymbol{\Gamma}\tilde{\boldsymbol{\theta}}\|_2^2\right] \\ &\propto \gamma_j \mathrm{logit}(\rho) + \gamma_j \mathrm{E}_{-\gamma_j}\left[\sigma^{-2}\left((\boldsymbol{y}-\boldsymbol{X}_{-j}\boldsymbol{\Gamma}_{-j}\tilde{\boldsymbol{\theta}}_{-j})^{\top}\boldsymbol{X}_j\tilde{\theta}_j - \frac{1}{2}\tilde{\theta}_j^2\boldsymbol{X}_j^{\top}\boldsymbol{X}_j\right)\right] \\ &\propto \gamma_j\left\{\mathrm{E}_q[\mathrm{logit}(\rho)] + \mathrm{E}_q[\sigma^{-2}]\left(\boldsymbol{y}^{\top}\boldsymbol{X}_j\mathrm{E}_q[\tilde{\theta}_j] - \boldsymbol{X}_j^{\top}\boldsymbol{X}_{-j}\mathrm{E}_q[\boldsymbol{\Gamma}_{-j}]\mathrm{E}_q[\tilde{\theta}_j\tilde{\boldsymbol{\theta}}_{-j}] - \frac{1}{2}\mathrm{E}_q[\tilde{\theta}_j^2]\boldsymbol{X}_j^{\top}\boldsymbol{X}_j\right)\right\}\end{aligned}$$

이에 따라 γ_j는 아래의 베르누이분포를 따른다.

$$\begin{aligned}&q^{\star}(\gamma_j) = Bernoulli(\omega_j), \\ &\omega_j = \mathrm{expit}\left\{\mathrm{E}_q[\mathrm{logit}(\rho)] + \mathrm{E}_q[\sigma^{-2}]\left(\boldsymbol{y}^{\top}\boldsymbol{X}_j\mathrm{E}_q[\tilde{\theta}_j] - \boldsymbol{X}_j^{\top}\boldsymbol{X}_{-j}\mathrm{E}_q[\boldsymbol{\Gamma}_{-j}]\mathrm{E}_q[\tilde{\theta}_j\tilde{\boldsymbol{\theta}}_{-j}] - \frac{1}{2}\mathrm{E}_q[\tilde{\theta}_j^2]\boldsymbol{X}_j^{\top}\boldsymbol{X}_j\right)\right\}\end{aligned}$$

다른 변분분포 계산에 필요한 기댓값 계산은 다음과 같다.

$$\begin{aligned}\mathrm{E}_q[\boldsymbol{\gamma}] &= (\omega_1,\ldots,\omega_p)^{\top} = \boldsymbol{\omega} \\ \mathrm{E}_q[\boldsymbol{\Gamma}] &= diag(\boldsymbol{\omega}) = \boldsymbol{\Omega} \\ \mathrm{E}_q[\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}}] &= \boldsymbol{\omega}\boldsymbol{\omega}^{\top} + \boldsymbol{\Omega}\odot(\mathrm{I}-\boldsymbol{\Omega})\end{aligned}$$

7.2. 평균장 변분추론 알고리즘

수렴조건을 만족할 때까지 다음 과정을 반복한다.

FOR i=1,2,...

Step 1) 변분분포 $q_{[i]}(\alpha)$ 갱신

Step 2) 변분분포 $q_{[i]}(\tilde{\boldsymbol{\theta}})$ 갱신

Step 3) 변분분포 $q_{[i]}(\tau_1^2),\ldots,q_{[i]}(\tau_p^2)$ 갱신

Step 4) 변분분포 $q_{[i]}(\rho)$ 갱신

Step 5) 변분분포 $q_{[i]}(\sigma^2)$ 갱신

Step 6) 변분분포 $q_{[i]}(\gamma_1),\ldots,q_{[i]}(\gamma_p)$ 갱신

수렴조건은 Huang et al. (2016)과 Ray and Szabó (2018)을 따라 다음과 같이 정의한다.

$$\begin{aligned}&\Delta_H = \max_{j=1,\ldots,p}|H(\mathrm{E}_q[\gamma_{[i],j}]) - H(\mathrm{E}_q[\gamma_{[i-1],j}])| \le 0.001, \\ &\text{where } H(p) = p\log p - (1-p)\log(1-p)\end{aligned}$$

$\mathrm{E}_q[\gamma_{[i]}^j]$은 $q_{[i]}(\boldsymbol{\gamma})$의 기댓값에 해당한다.

7.3. 구조화된 변분분포 $q(\boldsymbol{\gamma})$의 유도

$$\begin{aligned}\log q^{\star}(\boldsymbol{\gamma}) &\propto \mathrm{E}_{-\boldsymbol{\gamma}}\left[\sum_{j=1}^{p}\gamma_j\mathrm{logit}(\rho) - \frac{1}{2\sigma^2}\|(\boldsymbol{y}-\alpha\mathbf{1}_n) - \boldsymbol{X}\boldsymbol{\Gamma}\tilde{\boldsymbol{\theta}}\|_2^2\right] \\ &\propto \mathrm{E}_{-\boldsymbol{\gamma}}\left[\sum_{j=1}^{p}\gamma_j\left\{\mathrm{logit}(\rho) + \sigma^{-2}\tilde{\theta}_j(\boldsymbol{y}-\alpha\mathbf{1}_n)^{\top}\boldsymbol{X}_j\right\} - \frac{1}{2\sigma^2}\left(\sum_{j=1}^{p}\gamma_j\tilde{\theta}_j\boldsymbol{X}_j^{\top}\right)\left(\sum_{j=1}^{p}\gamma_j\tilde{\theta}_j\boldsymbol{X}\right)\right] \\ &\propto \boldsymbol{\psi}^{\top}\boldsymbol{\gamma} + \boldsymbol{\gamma}^{\top}\boldsymbol{\Psi}\boldsymbol{\gamma}\end{aligned}$$

$\boldsymbol{\psi}$와 $\boldsymbol{\Psi}$의 각 요소는 다음과 같다.

$$\begin{aligned}\psi_j &= \mathrm{E}_q[\mathrm{logit}(\rho)] + \mathrm{E}_q[\sigma^{-2}]\mathrm{E}_q[\tilde{\theta}_j](\boldsymbol{y}-\mathrm{E}_q[\alpha]\mathbf{1}_n)^{\top}\boldsymbol{X}_j \\ \Psi_{ij} &= -\frac{\mathrm{E}_q[\sigma^{-2}]}{2}\left(\mathrm{E}_q[\tilde{\theta}_i\tilde{\theta}_j]\boldsymbol{X}_i^{\top}\boldsymbol{X}_j\right)\end{aligned}$$

다른 변분분포 계산에 필요한 기댓값 계산은 다음과 같다.

$$\mathrm{E}_q[\gamma_j] = \sum_{\boldsymbol{\gamma}_{-j}} q(\gamma_j = 1, \boldsymbol{\gamma}_{-j})$$
$$\mathrm{E}_q[\gamma_i \gamma_j] = \sum_{\boldsymbol{\gamma}_{-ij}} q(\gamma_i = 1, \gamma_j = 1, \boldsymbol{\gamma}_{-ij})$$

7.4. SVI-S 알고리즘

수렴조건을 만족할 때까지 다음 과정을 반복한다.

FOR i=1,2,...

Step 1) 변분분포 $q_{[i]}(\alpha)$ 갱신

Step 2) 변분분포 $q_{[i]}(\tilde{\boldsymbol{\theta}})$ 갱신

Step 3) 변분분포 $q_{[i]}(\tau_1^2), \ldots, q_{[i]}(\tau_p^2)$ 갱신

Step 4) 변분분포 $q_{[i]}(\rho)$ 갱신

Step 5) 변분분포 $q_{[i]}(\sigma^2)$ 갱신

Step 6) - FOR $t = 1 : T-1$,

- IF $\mathrm{ESS}_t < \mathrm{Threshold}$:

샘플들 $\{\boldsymbol{\gamma}_{[i],t}^{(s)}\}_{s=1}^{S}$의 가중치 $\{W_{[i],t}^{(s)}\}_{s=1}^{S}$를 계산하고 normalize시킨 후 이를 바탕으로 resampling을 한다.

- 새로운 샘플들 $\{\boldsymbol{\gamma}_{[i],t+1}^{(s)}\}_{s=1}^{S}$을 중간분포 $q_{[i],t}(\boldsymbol{\gamma};\xi_i) \propto \tilde{q}_{[i],t}(\boldsymbol{\gamma})^{\xi_i}$에 불변한 전이함수 $K_{[i],t}$를 사용하여 샘플링한다, i.e., $\boldsymbol{\gamma}_{[i],t+1}^{(s)} \sim K_{[i],t}(\cdot \mid \boldsymbol{\gamma}_{[i],t}^{(s)})$

- 최종 가중치 $\{W_{[i],T}^{(s)}\}_{s=1}^{S}$로 $q_{[i]}(\boldsymbol{\gamma};\xi_i)$에 대한 기댓값을 근사한다.

Step 7) ξ_{i+1} 갱신

- IF $\xi_i < 1$: $\xi_{i+1} = \xi_i + \eta$
- ELSE : $\xi_{i+1} = 1$

수렴조건은 평균장 변분추론 알고리즘의 수렴조건과 동일하며 모든 실험에서 η값을 동일하게 0.1로 정하였다.

8. 참고문헌

- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association, 112*, 859-877
- Carbonetto, P., and Stephens, M. (2012). Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis, 7*, 73-107
- Carbonetto, P., Schmidt-Hieber, J., and van der Varrt, A. (2015). Bayesian Linear Regression With Sparse Priors. *The Annals of Statistics, 43*, 1986-2018
- Chenguang, D., Heng, J., Jacob, P. E., and Whiteley, N. (2022). An Invitation to Sequential Monte Carlo Samplers. *Journal of the American Statistical Association, 117*, 1587-1600
- Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo Samplers.

Journal of the Royal Statistical Society, Series B, 68, 411-436
- Hoffman, M. D., and Blei, D. M. (2015). Structured stochastic variational inference. In *Artificial Intelligence and Statistics, 38*, 361-369
- Huang, X., Wang, J., and Liang, F. (2016). A Variational Algorithm for Bayesian Variable Selection. *arXiv:1602.07640*
- Katahira, K., Watanabe, K., and Okada, M. (2007). Deterministic Annealing Variant of Variational Bayes Method. *Journal of Physics: Conference Series, 95*(1), 012015
- Mandt, S., McInerney, J., Abrol, F., Ranganath, R., and Blei, D. (2016). Variational Tempering. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*.
- Mimno, D., Hoffman, M. D., and Blei, D. M. (2012). Sparse Stochastic Inference for Latent Dirichlet Allocation. In *Proceedings of the 29th International Conference on Machine Learning*.
- Mitchell, T. J., and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. J*ournal of the American Statistical Association, 83*, 1023-1032
- Neal, R. M. (2001). Annealed Importance Sampling. *Statistics and Computing, 11*, 125-139
- Ormerod, J. T., You, C., and Muller, S. (2017). A Variational Bayes Approach to Variable Selection. *Electronic Journal of Statistics, 11*, 3549-3594
- Ray, K., and Szabo, B. (2022). Variational Bayes for High-Dimensional Linear Regression With Sparse Priors. *Journal of the American Statistical Association, 117*, 1270-1281
- Ročková, V., and George, E. I. (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association, 109*, 828-84
- Ročková, V., and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association, 113*, 431-44