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

Jinwoo Lee [1], Seonghyun Jeong[1,2], and Taeyoung Park[1,2]

[1]Department of Statistics and Data Science, Yonsei University; [2]Department of Applied Statistics, Yonsei University

1. Introduction

1.1. Bayesian Variable Selection

- Consider variable selection for linear regression with continuous outcomes,

$$y = \alpha + \sum_{j=1}^{p} X_j\theta_j + \epsilon,\ \epsilon \sim N(0, \sigma^2) \tag{1}$$

- Bayesian model selection can be modeled with a spike and slab distribution,

$$\begin{aligned} \pi(\theta_j|\gamma_j) &= \gamma_j\pi(\tilde{\theta}_j) + (1-\gamma_j)\delta_0, \qquad (2)\\ \pi(\gamma_j) &= \text{Bernoulli}(\rho). \qquad (3)\end{aligned}$$

- To control sparsity, $\pi(\tilde{\theta}_j) = \text{Laplace}(\tilde{\theta}_j; \lambda = 0.001)$ is set to be a heavy tailed distribution and $\rho \sim \text{Beta}(1, p)$ (Ročková et al., 2014, Ray et al., 2021).
- Since α and σ^2 are common to all models, an improper prior is specified as $\pi(\alpha, \sigma^2) \propto 1/\sigma^2$.
- A posterior distribution of $\mathbf{h} = (\tilde{\boldsymbol{\theta}}, \boldsymbol{\gamma}, \rho, \alpha, \sigma^2)$ cannot be obtained in a closed form.
- MCMC is available but computationally inefficient with high-dimensional variables.

1.2. Variational Inference (VI)

- Finds an approximate posterior by minimizing the Kullback-Leibler divergence, which is equivalent to maximizing the evidence lower bound (ELBO):

$$q(\mathbf{z}) = \arg\max_{q\in Q} \underbrace{\mathbb{E}_{q(\mathbf{h})}[\log p(\mathbf{h}, \mathbf{y}) - \log q(\mathbf{h})]}_{\text{ELBO}}, \tag{4}$$

where distributions in the mean field family Q are fully factorized, i.e., $q(\mathbf{h}) = \prod_j q_j(h_j)$.

- The ELBO objective is more vulnerable to nonconvexity because of the independence assumption on $\mathbf{h}$.
- Initialization and update-ordering become issues when dealing with multiple local optima.

1.3. Structured Variational Inference (SVI)

- Restoring some dependencies results in the structured variational inference (SVI).
- SVI not only allows for more accurate approximation of the exact posterior, but it is also more resistant to the initialization problem.
- However, SVI has limitations because a structured distribution must be derived in a closed form.

2. Motivation

2.1. Current Methods

- Previous methods assume a variational distribution of $\boldsymbol{\gamma}$ is fully factorized, i.e., $q(\boldsymbol{\gamma}) = \prod_{j=1}^{p} q_j(\gamma_j)$ (Carbonetto et al., 2012, Huang et al., 2016, Ormerod et al., 2017, and Ray et al., 2021).
- While the assumption of complete independence makes the derivation of each $q_j(\gamma_j)$ tractable, the discrepancy from the exact posterior becomes severe for high-dimensional covariates.
- They mitigate the local optimum issue by identifying a more sensible initial point or by establishing update ordering rules.
- We derive $q(\boldsymbol{\gamma})$ without the assumption of mutual independence for a more precise approximation and use deterministic annealing to solve the initial point problem.

2.2. Challenges

- A variational distribution for $\boldsymbol{\gamma}$ can be derived in a quadratic form,

$$q(\boldsymbol{\gamma}) = \frac{\tilde{q}(\boldsymbol{\gamma})}{Z}, \quad \tilde{q}(\boldsymbol{\gamma}) = \exp(\boldsymbol{\psi}^\top\boldsymbol{\gamma} + \boldsymbol{\gamma}^\top\boldsymbol{\Psi}\boldsymbol{\gamma}), \quad Z = \sum_{\boldsymbol{\gamma}} \tilde{q}(\boldsymbol{\gamma}). \tag{5}$$

- To perform VI, we have to evaluate $\mathbb{E}_{q(\boldsymbol{\gamma})}[\cdot]$ for every VI iteration, resulting in an infeasible amount of combinatorial sum over 2^p terms,
- Mimno et al. (2012) used Gibbs sampling to approximate the expectation.
- Gibbs sampling is susceptible to producing suboptimal results for complex distributions, such as multimodality and highly correlated variables, resulting in error propagation at each iteration step of VI.
- We propose to approximate the expectation with the sequential Monte Carlo (SMC, Del Moral et al., 2006) sampler nested in a variational inference scheme.

3. Methodology

3.1. Background

- The SMC sampler is an importance sampling (IS) technique to approximate an expectation with three main ingredients: intermediate distributions, resampling, and transition.
- **IS** approximates $\mathbb{E}_{q(\boldsymbol{\gamma})}[f(\boldsymbol{\gamma})]$ with weights $w^{(1)}, \ldots, w^{(S)}$ using particles $\boldsymbol{\gamma}^{(s)} \overset{iid}{\sim} \eta(\boldsymbol{\gamma})$, $s = 1, \ldots, S$, i.e.,

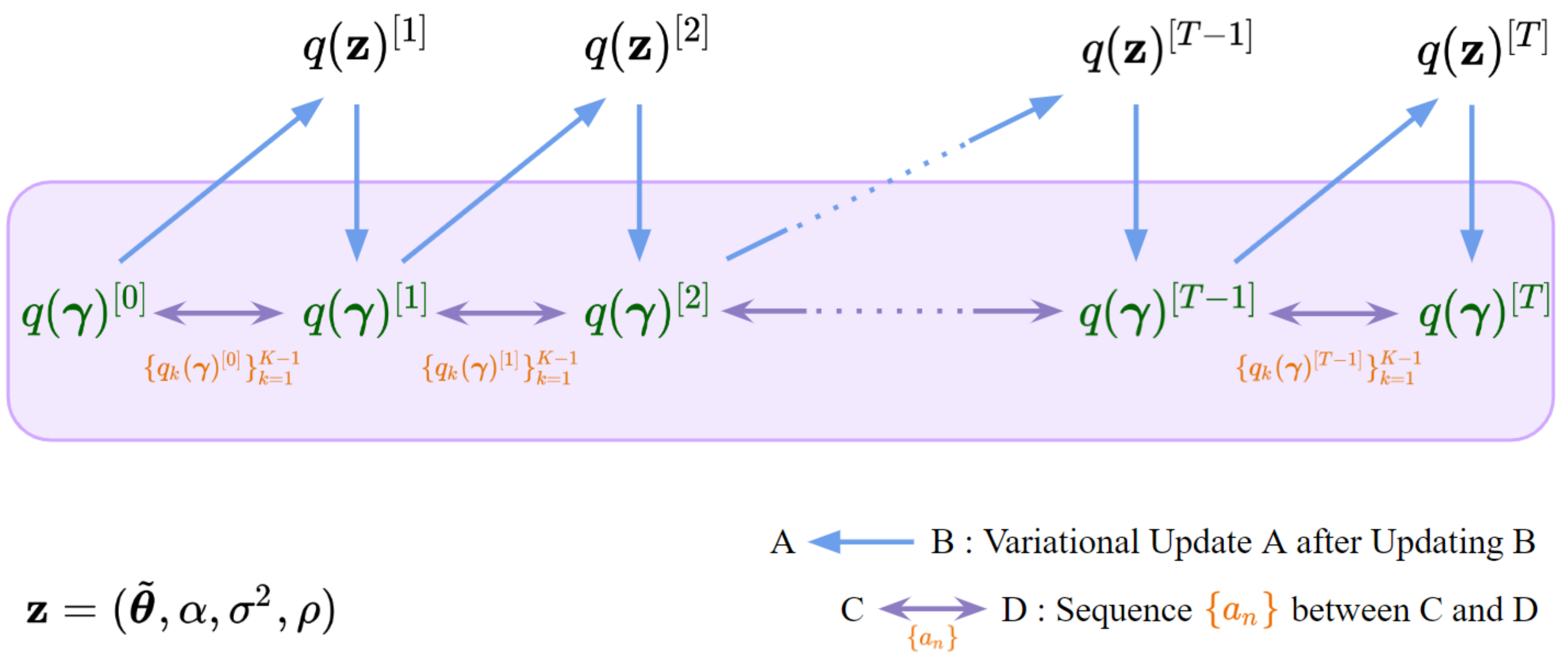
$$\mathbb{E}_{q(\boldsymbol{\gamma})}[f(\boldsymbol{\gamma})] \approx \sum_{s=1}^{S} w^{(s)} f(\boldsymbol{\gamma}^{(s)}), \quad w^{(s)} = \frac{W^{(s)}}{\sum_{s'} W^{(s')}}, \quad W^{(s)} = \frac{q(\boldsymbol{\gamma}^{(s)})}{\eta(\boldsymbol{\gamma}^{(s)})}, \tag{6}$$

but the quality of the approximation degrades as the difference between $q(\boldsymbol{\gamma})$ and $\eta(\boldsymbol{\gamma})$ increases.

- **Intermediate distributions** are constructed to overcome the difference by bridging smoothly between $q(\boldsymbol{\gamma})$ and $\eta(\boldsymbol{\gamma})$.
- **Resampling** helps to utilize only promising particles by duplicating particles with high weights and discarding low weight particles when the effective sample size (ESS) of particles is below the prespecified threshold.
- **Transition** allows particles from $\eta(\boldsymbol{\gamma})$ to move to high density region of $q(\boldsymbol{\gamma})$ via sequence of kernel functions invariant to intermediate distributions with resampling.

3.2. Nested SMC sampler

- Our proposition is to use two levels of intermediate distributions.
- Assume that $q(\boldsymbol{\gamma})^{[T]}$ is a converged variational distribution after the number T of VI iterations, and $q(\boldsymbol{\gamma})^{[0]}$ is an initial distribution.
- The 1st level of intermediate distributions are variational distributions before $q(\boldsymbol{\gamma})^{[T]}$, i.e., $\{q(\boldsymbol{\gamma})^{[t]}\}_{t=1}^{T-1}$
- The 2nd level distributions are annealed distributions $\{q_k(\boldsymbol{\gamma})^{[t]}\}_{k=1}^{K-1}$ that bridge between $q(\boldsymbol{\gamma})^{[t]}$ and $q(\boldsymbol{\gamma})^{[t+1]}$, which are defined as $q_k(\boldsymbol{\gamma})^{[t]} \propto (q(\boldsymbol{\gamma})^{[t]})^{1-\beta_k}(q(\boldsymbol{\gamma})^{[t+1]})^{\beta_k}$ with a sequence of parameters $0 = \beta_0 < \cdots < \beta_K = 1$.



$q(\mathbf{z})^{[1]}$ $q(\mathbf{z})^{[2]}$ $q(\mathbf{z})^{[T-1]}$ $q(\mathbf{z})^{[T]}$

$q(\boldsymbol{\gamma})^{[0]}$ $q(\boldsymbol{\gamma})^{[1]}$ $q(\boldsymbol{\gamma})^{[2]}$ $q(\boldsymbol{\gamma})^{[T-1]}$ $q(\boldsymbol{\gamma})^{[T]}$

$\{q_k(\boldsymbol{\gamma})^{[0]}\}_{k=1}^{K-1}$ $\{q_k(\boldsymbol{\gamma})^{[1]}\}_{k=1}^{K-1}$ $\{q_k(\boldsymbol{\gamma})^{[T-1]}\}_{k=1}^{K-1}$

A ← B : Variational Update A after Updating B

$\mathbf{z} = (\tilde{\boldsymbol{\theta}}, \alpha, \sigma^2, \rho)$

C $\overset{\{a_n\}}{\longleftrightarrow}$ D : Sequence $\{a_n\}$ between C and D

Figure 1: Illustration of SMC sampler nested in VI iterations.

- A Markov chain transition $T_{[t],k}(x'|x)$ invariant to $q_k(\boldsymbol{\gamma})^{[t]}$ represents a probability of moving from x to x'.
- To produce an importance weight of the particle $\boldsymbol{\gamma}_{[t],k}$ generated from the kth annealed distribution between $q(\boldsymbol{\gamma})^{[t]}$ and $q(\boldsymbol{\gamma})^{[t+1]}$, a sequence of particles is generated as follows:

$$\boldsymbol{\gamma}_{[0],0} \to \boldsymbol{\gamma}_{[0],1} \to \cdots \to \boldsymbol{\gamma}_{[0],K-1} \to \boldsymbol{\gamma}_{[1],0} \to \boldsymbol{\gamma}_{[1],1} \to \cdots \to \boldsymbol{\gamma}_{[t],k-1} \to \boldsymbol{\gamma}_{[t],k} \tag{7}$$

where $\boldsymbol{\gamma}_{[0],0} \sim q(\boldsymbol{\gamma})^{[0]}$ and $\boldsymbol{\gamma}_{[t],k-1} \to \boldsymbol{\gamma}_{[t],k}$ implies $\boldsymbol{\gamma}_{[t],k} \sim T_{[t],k}(\cdot|\boldsymbol{\gamma}_{[t],k-1})$.

- Then the weight $W_{[t],k+1}$ can be computed inductively, i.e.,

$$W_{[0],k} = \prod_{l=0}^{k} \frac{\tilde{q}_{l+1}(\boldsymbol{\gamma}_{[0],l})^{[0]}}{\tilde{q}_l(\boldsymbol{\gamma}_{[0],l})^{[0]}},\ W_{[1],k} = W_{[0],K}\prod_{l=0}^{k} \frac{\tilde{q}_{l+1}(\boldsymbol{\gamma}_{[1],l})^{[1]}}{\tilde{q}_l(\boldsymbol{\gamma}_{[1],l})^{[1]}}, \cdots, W_{[t],k+1} = W_{[t-1],K}\prod_{l=0}^{k} \frac{\tilde{q}_{l+1}(\boldsymbol{\gamma}_{[t],l})^{[t]}}{\tilde{q}_t(\boldsymbol{\gamma}_{[t],l})^{[t]}} \tag{8}$$

- The weights $W_{[t],k+1}^{(1)}, \ldots, W_{[t],k+1}^{(S)}$ are normalized as $w_{[t],k+1}^{(s)} = W_{[t],k+1}^{(s)} / \sum_{s'=1}^{S} W_{[t],k+1}^{(s')}$ and used for resampling, i.e.,

$$\{\boldsymbol{\gamma}_{[t],k+1}^{(s)}\}_{s=1}^{S} = \begin{cases} \text{resample } \{\boldsymbol{\gamma}_{[t],k}^{(s)}\}_{s=1}^{S} \text{ with } \{w_{[t],k+1}^{(s)}\}_{s=1}^{S} & \text{if ESS} < S/2 \\ \{\boldsymbol{\gamma}_{[t],k}^{(s)}\}_{s=1}^{S} & \text{otherwise} \end{cases} \tag{9}$$

when $k \le K-2$ and the expectation $\mathbb{E}_{q(\boldsymbol{\gamma})^{[t]}}[f(\boldsymbol{\gamma})] \approx \sum_{s=1}^{S} w_{[t],K}^{(s)} f(\boldsymbol{\gamma}_{[t],K-1}^{(s)})$ is approximated when $k = K-1$.

4. Experiments

- Assume there are n observations and p variables with s nonzero effects and $p - s$ zero effects.
- The nonzero coefficients are uniformly sampled in $\theta \in [-10, -1] \cup [1, 10]$.
- We evaluate the performance of model selection with two indexes: false negative number of linear covariates misclassified as zero effect FN $= \sum_{j=1}^{p} \mathbb{1}(\mathbb{E}_{q(\boldsymbol{\gamma})}(\gamma_j|\mathbf{y}) < 0.5, \gamma_j = 1)$ and false positive number of noneffect covariates misclassified as linear effect FP $= \sum_{j=1}^{p} \mathbb{1}(\mathbb{E}_{q(\boldsymbol{\gamma})}(\gamma_j|\mathbf{y}) \ge 0.5, \gamma_j = 0)$.
- For the nested SMC sampler, we used 100 particles, 300 annealed distributions for each VI iteration, and the Gibbs sampler for particle transition.
- The proposed SVI with the nested SMC sampler (SVI-S) is compared with other methods: EMVS (Ročková et al, 2014), SSLASSO (Ročková et al, 2018), varbvs (Carbonetto et al, 2012), sparsevb (Ray et al, 2021), MFVI, SVI-G (Mimno et al., 2012).

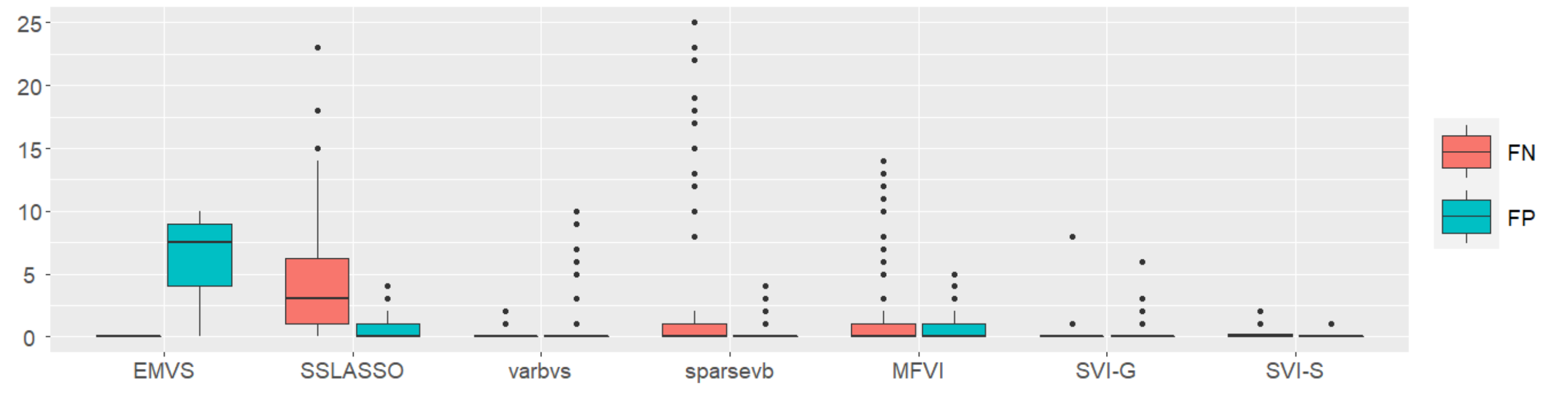


Figure 2: Mean FN and FP for 100 random datasets with setup $n = 50, p = 100, s = 10, \phi = 0.3$.

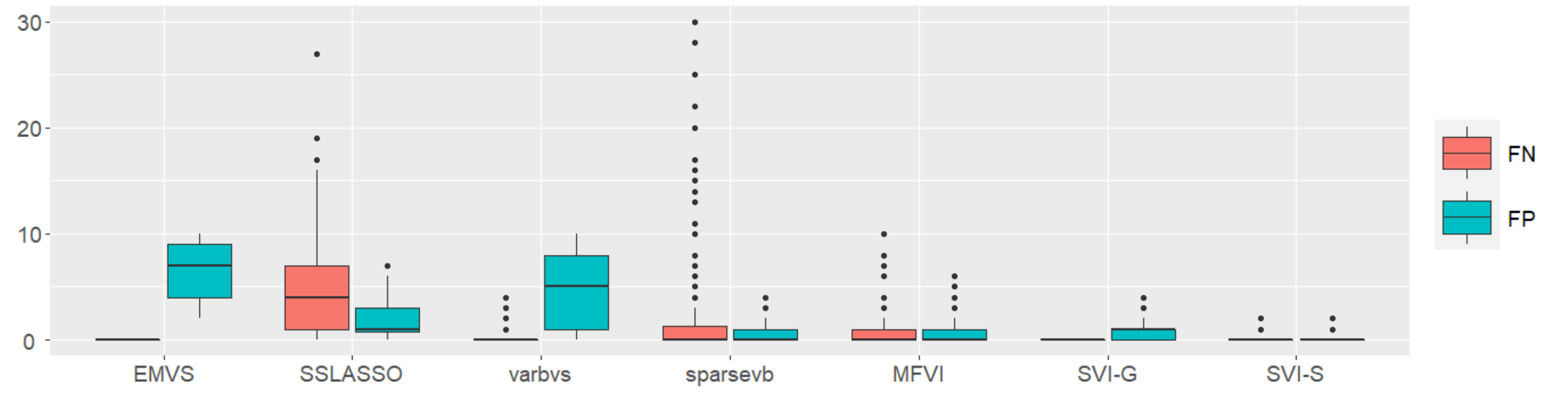


Figure 3: Mean FN and FP for 100 random datasets with setup $n = 50, p = 100, s = 10, \phi = 0.6$.

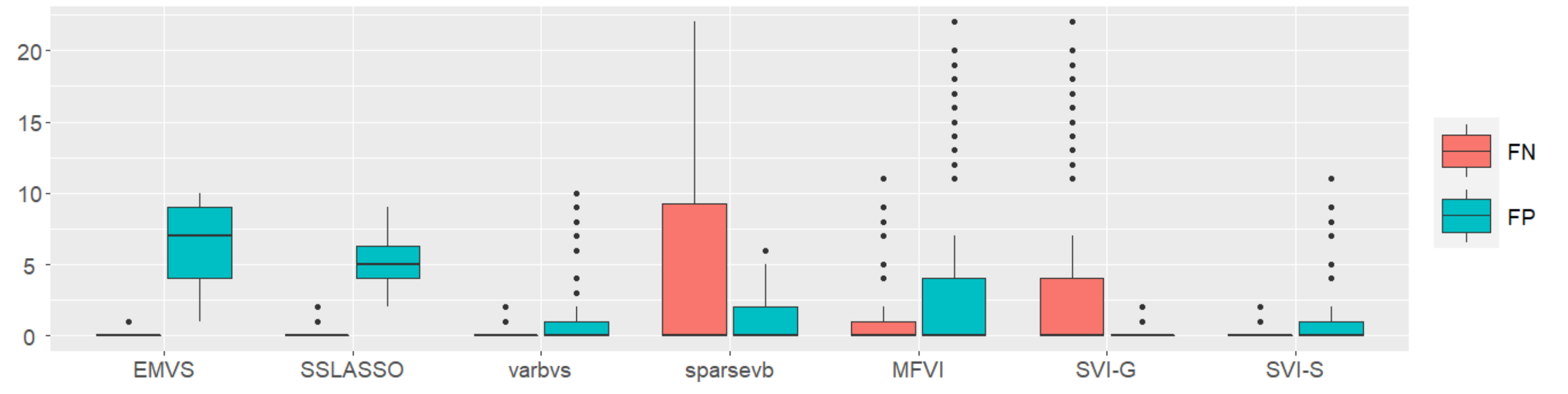


Figure 4: Mean FN and FP for 100 random datasets with setup $n = 50, p = 200, s = 10, \phi = 0.3$.

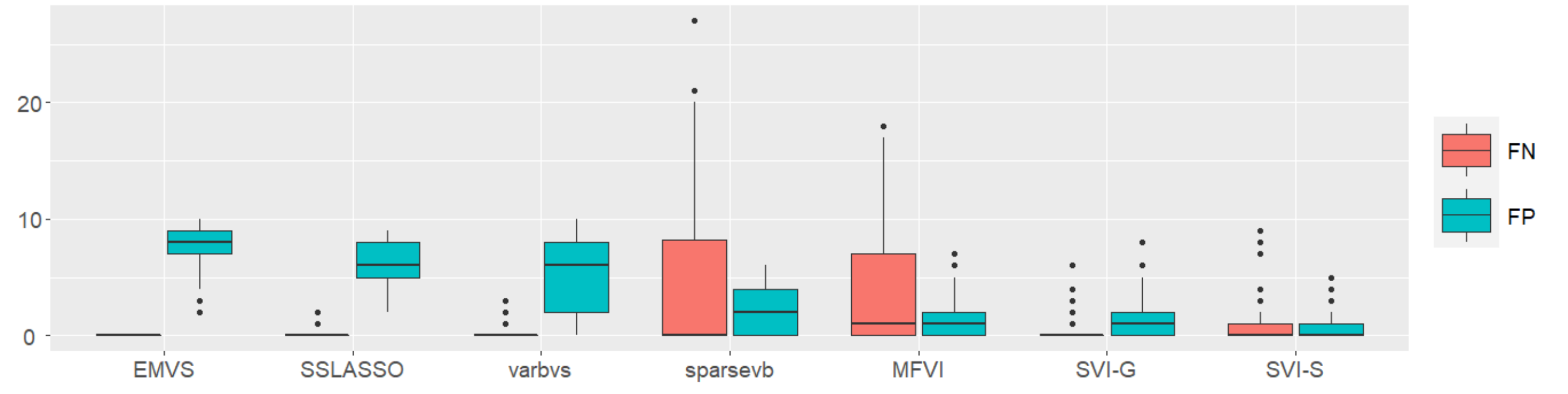


Figure 5: Mean FN and FP for 100 random datasets with setup $n = 50, p = 200, s = 10, \phi = 0.6$.

- SVI-S consistently produces good results for both FN and FP, compared to other methods.

5. Discussion

- Our work is currently in progress; we are developing a robust and faster method to deal with more high dimensional and correlated variables via other transition methods.