SuSiE: A Bayesian Method for Variable Selection in Regression

Method Paper Presentation

Rohan Dekate

Carlos Copana

Elaine Hu

Variable Selection: Challenges & Solutions

- **Challenge:** Scientific conclusions depend on **which variables are selected**, yet selection is hard when predictors are **strongly correlated**.
- **Problem Statement:** How can we **quantify uncertainty** in variable selection under high correlation?
- **Goal:** Identify **causal variants** that truly affect traits — aiming for **scientific insight**, not just predictive accuracy.

BVSR (Bayesian Variable Selection in Regression)

- Can assess uncertainty even with correlated variables.
- **Computationally intensive** and yields **complex posteriors**.

SuSiE (Sum of Single Effects):

- Builds on BVSR but is **faster, simpler, and more interpretable**.
- Uses **Iterative Bayesian Stepwise Selection (IBSS)** and **variational approximation** for efficient model fitting.
- Provides **credible sets** — groups of variables that express uncertainty when multiple correlated candidates compete.

Motivating Toy Example

Consider a linear model:

$$y = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

with predictors $x_1 = x_2$ and $x_3 = x_4$ and $\beta_1 \neq 0, \quad \beta_4 \neq 0, \quad \beta_2 = \beta_3 = 0.$

- We wish to infer the statement: $\beta_1 \neq 0$ **or** $\beta_2 \neq 0$ **and** $\beta_3 \neq 0$ **or** $\beta_4 \neq 0$.

Existing methods fall short:

- **LASSO / Elastic Net:** Picks one "best" model $\Rightarrow$ ignores uncertainty.
- **BVSR: Sparse priors** on β capture uncertainty; but **PIPs** lose detail.

Single-Effect Regression (SER): Only **one variable** has a **non-zero effect**.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad \mathbf{b} = b\boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}), \quad b \sim \mathcal{N}(0, \sigma_0^2)$$
$$\boldsymbol{\gamma} \mid \mathbf{X}, \mathbf{y} \sim \text{Mult}(1, \boldsymbol{\alpha}), \quad b \mid \mathbf{X}, \mathbf{y}, \gamma_j = 1 \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

- $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_p)$ are the prior inclusion probabilities.
- $\boldsymbol{\gamma} \in \{0, 1\}^p$ is p-vector of indicator variables.
- The single-effect vector $\mathbf{b}$ has exactly one non-zero element.

Sum of Single-Effects Regression Model: SuSiE

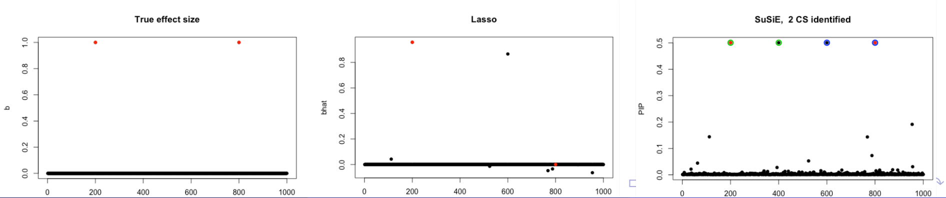
- **Idea:** Extend the SER model to allow for **multiple effects**.
- **Model:**

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

- Represent the regression coefficient vector as a **sum of single effects:**

$$\beta = \sum_{l=1}^{L} \mathbf{b}_l, \quad \mathbf{b}_l = b_l \gamma_l, \quad b_l \sim \mathcal{N}(0, \sigma_{0l}^2), \quad \gamma_l \sim \mathrm{Mult}(1, \boldsymbol{\pi}).$$

- When $L = 1$, SuSiE reduces to SER; when $L \ll p$, it approximates BVSR.
- Each component $\mathbf{b}_l$ corresponds to one **independent signal** in the data.



Iterative Bayesian Stepwise Selection (IBSS)

Goal: Efficiently fit the SuSiE model using repeated SER updates.

Algorithm outline:

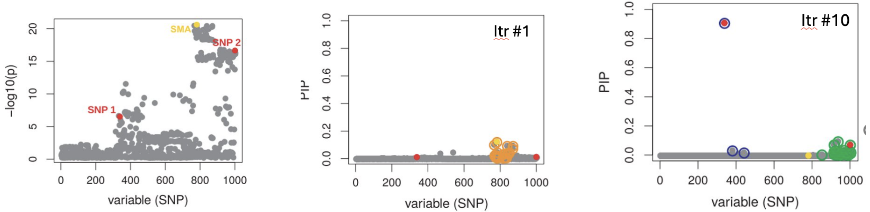
1. Initialize posterior means $\bar{\mathbf{b}}_l = 0$ for $l = 1, \ldots, L$.
2. Repeat until convergence:
    1. For each l:
        - Compute residuals excluding effect l: $\mathbf{r}_l = \mathbf{y} - \mathbf{X}\sum_{l' \neq l} \bar{\mathbf{b}}_{l'}$.
        - Fit SER to $\mathbf{r}_l$ and obtain : $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l^2) = SER(\mathbf{X}, \mathbf{r}_l; \sigma^2, \sigma_{0l}^2)$.
        - Update effect estimate: $\bar{\mathbf{b}}_l = \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_l$ (element-wise product).

Posterior Inference and Key features:

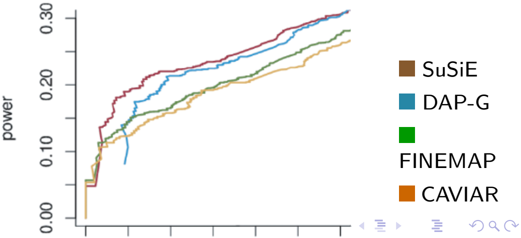
- Under SuSiE we obtain: $\mathrm{PIP}_j = \Pr(b^j \neq 0 \mid \mathbf{X}, \mathbf{y}) \approx 1 - \prod_{l \in \mathcal{L}} (1 - \alpha_{lj})$
- **Choice of L:** If L exceeds the number of detectable effects in the data, then many of the L credible sets are large.
- Identifiability and label switching $p(\mathbf{b}_1, \cdots, \mathbf{b}_L | \mathbf{y}) = p(\mathbf{b}_{\nu(1)}, \cdots, \mathbf{b}_{\nu(L)} | \mathbf{y})$

Simulation and Numerical Comparison

Setup: 6000 simulated datasets comparing SuSiE, DAP-G, FINEMAP, and CAVIAR on credible sets, runtime, and power–FDR trade-off.



Alg.	Time (s)	Credible Sets.
SuSiE	0.64	High power
DAP-G	2.87	Equal or less
FINEMAP	23.0	—
CAVIAR	2907	—



Limitation and Change Point Detection

Main Limitation:

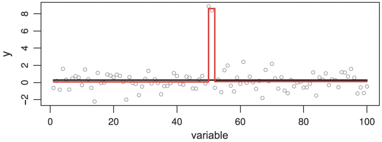
- IBSS uses coordinate-wise updates and can get stuck in **local optima**.
- Struggles to detect signals requiring **joint selection** of multiple correlated effects.

Example – Change Point Detection:

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \qquad \mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

where $\mathbf{X}$ encodes step functions marking potential change points.

- SuSiE reformulates this as regression, but IBSS may miss **adjacent change points** when signals cancel each other.



Future Directions:

- *Better Initialization; Optimization; Model extension*

References

- G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens (2020). *A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society, Series B*, **82**(5), 1273–1300. https://doi.org/10.1111/rssb.12388