THE UNIVERSITY OF CHICAGO

AN EXTENSION OF SUSIE MODEL WITH MIXTURE-GAUSSIAN PRIOR

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES DIVISION
IN CANDIDACY FOR THE DEGREE OF
MASTER

DEPARTMENT OF DEPARTMENT OF STATISTICS

BY
ZHENGYANG FANG

CHICAGO, ILLINOIS
SUMMER 2019

To my girlfriend Jian

without whom this thesis would have been completed earlier

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This thesis is under the supervision of Prof. Matthew Stephens. Thanks to Peter Carbonetto for the substantial instruction and advice. We also thank John Blischak for the development of the version control tool *workflowr*.

ABSTRACT

We extend the *Sum of Single Effects (SuSiE)* model - a Bayesian approach to variables selection in linear regression, to SuSiE-mixture model. SuSiE model assumes a slab-and-spike prior on the regression coefficients, and we generalize the prior to a mixture-Gaussian distribution centered at 0. We also extend the corresponding fitting procedure - Iterative Bayesian Stepwise Selection (IBSS) - which is a Bayesian analogue of stepwise selection methods. Specifically we introduce an additional ridge regression step to IBSS, based on variational approximation to the posterior distribution under the SuSiE-mixture model. Our methods provide extra flexibility to SuSiE model, and reduce the false-positive significantly. We demonstrate through simulated experiments.

CHAPTER 1

INTRODUCTION

Variable selection in linear regression has been an important problem for a long time, and it has a large range of methods and potential applications (Desboulets, 2018). The *Sum of Single Effect (SuSiE)* model (Wang et al., 2018) is a very recent work in Bayesian variable selection in linear regression. Motivated by genetic fine-mapping problems (Veyrieras et al., 2008; Schaid et al., 2018), the SuSiE model is particularly helpful when the explanatory variables are highly correlated, and the true regression coefficients are extremely sparse. Also, it provides a new and effective way to capture the uncertainty of the variable selection. The SuSiE model provides the *posterior inclusion probability (PIP)* for each variable, measuring the probability that a variable has a non-zero regression coefficient when the data is given.

However, in some cases, the highly sparsity assumption on the regression coefficients might be too strong. A more flexible assumption is, the majority of the regression coefficients is small but non-zero. In this paper, we extend the prior of the regression coefficients in the SuSiE model to this more flexible case. We call it the "SuSiE-mixture" model, because the marginal prior distribution of the regression coefficient is mixture-Gaussian, instead of "spike-and-slab" in the SuSiE model. Under appropriate settings, the "SuSiE-mixture" model outperforms the SuSiE model in variable selection.

The SuSiE model has a fast fitting procedure *Iterative Bayesian Stepwise Selection (IBSS)*, which computes the variational approximation to the posterior probability. We follow this idea and derive a variational inference algorithm for the SuSiE-mixture model.

The structure of the paper goes as follows. In **Chapter 2**, we introduce the SuSiE model and further details. In **Chapter 3**, we provide the formulation of our model and the fitting algorithm. In **Chapter 4**, we compare our model with the SuSiE model, and demonstrate the difference in the estimated PIP. In **Chapter 5** we end with a discussion of the limitations in our work.

CHAPTER 2

BACKGROUND

2.1 Introduction to the SuSiE model

The SuSiE model is an extension of the *single effect regression (SER)* model (Servin and Stephens, 2007). In the SER model we assume *exactly one of the p explanatory variables has a non-zero coefficient*. The SER model is easy to fit and provides useful inference result. Its further applications includes Pickrell (2014).

The SuSiE model extends the assumption of the SER model, assuming the effect to be *the sum of* a few "single effect" in the SER model . In other word, the SuSiE model assumes there are *at most L non-zero regression coefficients* in the linear regression model, where L is a pre-specified constant. Also we need $L << p$, where p is the number of the explanatory variables. Compared to the conventional *Bayesian Variable Selection Regression (BVSR)* (Mitchell and Beauchamp, 1988), the SuSiE model has a brand new modeling:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{2.1}$$

$$\mathbf{e} \sim N(0, \sigma^2 I_n) \tag{2.2}$$

$$\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}_l \tag{2.3}$$

$$\mathbf{b}_l = \boldsymbol{\gamma}_l b_l \tag{2.4}$$

$$\boldsymbol{\gamma}_l \sim Mult(1, \boldsymbol{\pi}) \tag{2.5}$$

$$b_l \sim N(0, \sigma_{0l}^2). \tag{2.6}$$

Here, $\mathbf{y}$ is the vector of response data with dimension n; $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ is an $n \times p$ data matrix with n observations of p explanatory variables; $\mathbf{e}$ is the independent error terms;

$\mathbf{b}$ is the p-vector of regression coefficients, it is the sum of L "single effects" $\mathbf{b}_1, \ldots, \mathbf{b}_L$; each of those "single effects" $\mathbf{b}_l$ has a indicator variable $\boldsymbol{\gamma}_l \in \{0, 1\}^p$, denoting which explanatory variable has the non-zero coefficient; also $b_1, \ldots, b_L$ are scalars representing the specific size of the "single effect"; $Mult(1, \boldsymbol{\pi})$ denotes the multinomial distribution on class counts obtained when only 1 sample is drawn with class probabilities given by $\boldsymbol{\pi}$, specifically it is a p-vector with only one element to be 1 and all the other elements to be 0; $\sigma_{01}^2, \ldots, \sigma_{0L}^2$ are the prior variances of the size of "single effects".

Also, we assume that $\mathbf{y}$ and the columns of $\mathbf{X}$ have been centered to have mean zero, so that we can avoid the intercept term. For the normal case where we don't have a specific preference for any of those explanatory variables in the prior, we can simply set $\boldsymbol{\pi} = (1/p, \ldots, 1/p)$.

2.2 Posterior inclusion probability

While the SuSiE model has a big range of applications, here we focus on the variable selection problem. Specifically we are interested in the marginal posterior inclusion probability

$$PIP_j := \mathbb{P}(b_j \neq 0 | \mathbf{X}, \mathbf{y}), 1 \leq j \leq p. \tag{2.7}$$

Where b_j is the j-th element of $\mathbf{b}$. The PIP measures how likely the j-th variable is a true factor given the data. It is very helpful in capturing the uncertainty in the variable selection procedure. The SuSiE model can compute the PIP effectively. In this paper, we compare the PIP with the true label of variables to measure the performance for different models. We will go through further details in the following chapters.

CHAPTER 3

THE SUSIE-MIXTURE MODEL

3.1 Generalize the prior assumption

In the SuSiE model, each of the L single effects $\mathbf{b}_1, \ldots, \mathbf{b}_L$ has *exactly one* non-zero coefficient out of the p possible variables. Let $\sum_{l=1}^{L} \mathbf{b}_l = \mathbf{b} = (\beta_1, \beta_2, \ldots, \beta_p)^T$. For a regression coefficient β_j $(1 \leq j \leq p)$, if it happens to be "the none-zero one" in some $\mathbf{b}_l$ where $1 \leq l \leq L$, then β_j will be non-zero. Otherwise β_j has to be zero. Thus the marginal prior distribution of β_j is spike-and-slab: it has a point mass at zero, and a curve density peaked at zero.

The SuSiE model assumes the coefficients of those irrelevant variables to be exactly zero. However, a more flexible assumption is, the irrelevant variables *may have small effects* on the response, *instead of having no effect*. And of course, the relevant variables have a strong influence on the response. In the original BVSR setting, we try to select any non-zero effects. While in this setting, the goal of variable selection is to find the variables with *a large effect*, and ignore those with *a small effect*. Specifically, we modify the point mass in the spike-and-slab prior, to a Gaussian distribution centered at zero with a very small variance σ_b^2. Then the prior of the regression coefficient becomes a mixture-Gaussian distribution. This assumption is a more general case (when $\sigma_b^2 = 0$, it degenerates to the original spike-and-slab prior) and it allows more flexibility (Zhou et al., 2013). In many settings it matches the real case better. Figure 3.1 intuitively shows the difference of the two prior distributions.

3.2 Model settings

We apply the new mixture-Gaussian prior to the SuSiE model, and get the formulation of the SuSiE-mixture model as follows

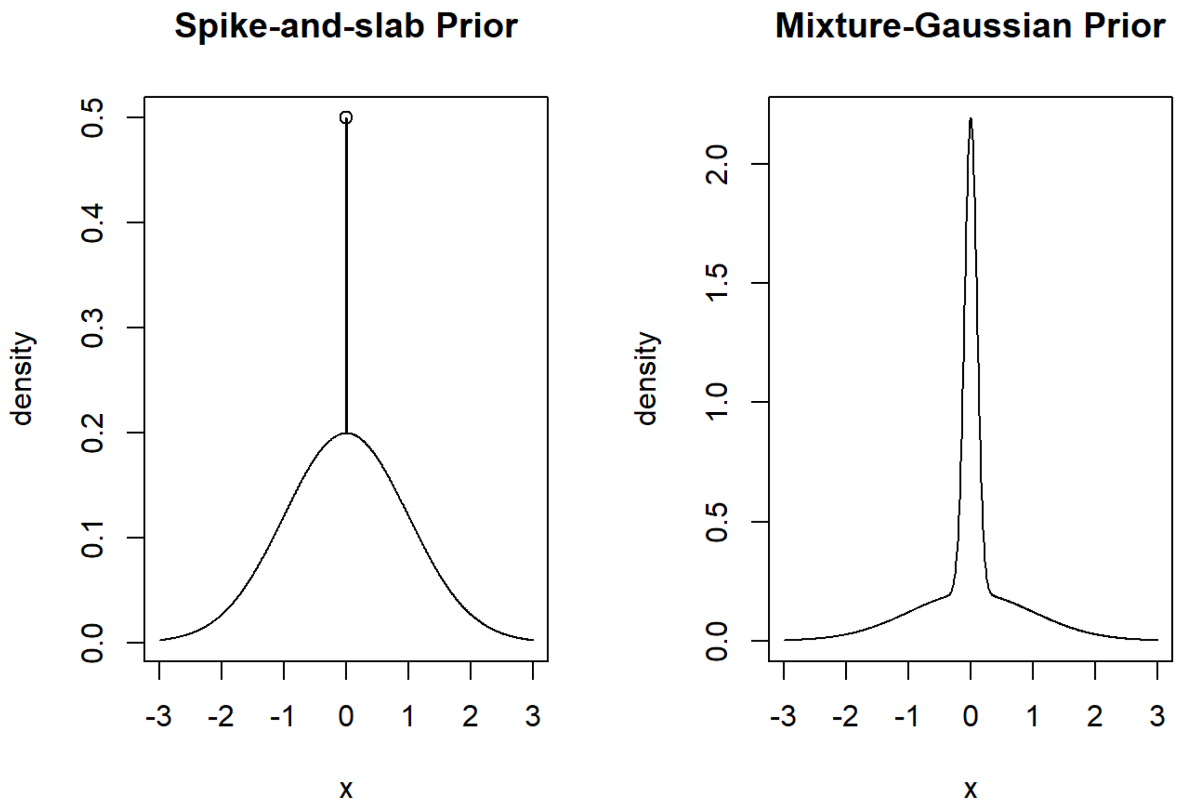

Figure 3.1: Different prior distributions of the regression coefficient

$$\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{X}\mathbf{b} + \mathbf{e} \tag{3.1}$$

$$\mathbf{e} \sim N(0, \sigma^2 I_n) \tag{3.2}$$

$$\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}_l \tag{3.3}$$

$$\mathbf{b}_0 \sim N(0, \sigma_b^2 I_p) \tag{3.4}$$

$$\mathbf{b}_l = \boldsymbol{\gamma}_l b_l, 1 \leq l \leq L \tag{3.5}$$

$$\boldsymbol{\gamma}_l \sim Mult(1, \boldsymbol{\pi}) \tag{3.6}$$

$$b_l \sim N(0, \sigma_{0l}^2). \tag{3.7}$$

The additional p-vector $\mathbf{b}_0$ represents the small effects for all variables. Those small effects are independent to each other, and share the same prior variance σ_b^2.

A notable fact is that, if we let $L = 0$, i.e. no "single effects", then the SuSiE-mixture model is equivalent to ridge regression. In fact, solving ridge regression plays an important role in fitting the SuSiE-mixture model.

3.3 Model fitting

The PIP is critical in handling the uncertainty in variable selection, in this paper we also focus on calculating the PIP for the SuSiE-mixture model. The definition of PIP is the same as in the SuSiE model (see equation 2.7). Although we introduce an additional term $\mathbf{b}_0$ to the model, as long as we estimate the posterior distribution of $(\mathbf{b}_1, \ldots, \mathbf{b}_L)$, $\mathbf{b}_0$ is no longer involved in computing the PIP.

We make an intuitive comparison between SuSiE and SuSiE-mixture. In the SuSiE-mixture model, some signals in the data will be attributed to the additional term $\mathbf{b}_0$, and can no longer be the evidence for those "single effects". Hence we will make fewer discoveries than the SuSiE model does. Also, the larger σ_b^2 is, the more signals $\mathbf{b}_0$ will be in charge of, the less "single effects" we can finally discover.

As we noted, the SuSiE-mixture model is closely related to ridge regression. Note that in the Bayesian form of ridge regression

$$
\begin{aligned}
&\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\
&\boldsymbol{\beta} \sim N(0, \sigma_b^2 I_p), \\
&\boldsymbol{\epsilon} \sim N(0, I_n).
\end{aligned}
$$

Compare with the optimization form of ridge regression

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

When the tuning parameter $\lambda = \sigma^2/\sigma_b^2$, the estimated $\hat{\boldsymbol{\beta}}$ are exactly the same for both forms. If the tuning parameter λ is known, i.e. (σ^2/σ_b^2) is known, then fitting ridge regression will be simple. Analogously, in the SuSiE-mixture model, if we know the value of σ_b^2/σ^2, the fitting procedure will be easy and neat.

3.3.1 Algorithm for known σ_b^2/σ^2

In this algorithm, we need the function for Choleski decomposition: given a positive semi-definite matrix S, return a lower triangular matrix L s.t. $LL^T = S$.

We also need the SuSiE function: given the data $\mathbf{X}, \mathbf{y}$, and the number of "single effects" L, return the PIP. See the detail in Wang et al. (2018).

Algorithm 1 Fit SuSiE-mixture for known σ_b^2/σ^2

```
**Given** $\mathbf{X}, \mathbf{y}, L, r = \sigma_b^2/\sigma^2$
$S \leftarrow r\mathbf{X}\mathbf{X}^T + I_n$
$L \leftarrow Cholesky(S)$
$\tilde{\mathbf{X}} \leftarrow L^{-1}\mathbf{X}$
$\tilde{\mathbf{y}} \leftarrow L^{-1}\mathbf{y}$
$PIP \leftarrow susie(\tilde{\mathbf{X}}, \tilde{\mathbf{y}}, L)$
**Return** $PIP$.
```

See the derivation in **Appendix A**.

3.3.2 For unknown σ_b^2/σ^2

In most practical cases, the ratio σ_b^2/σ^2 is unknown, and we need to estimate it. We modify the algorithm for fitting SuSiE - *Iterative Bayesian Stepwise Selection (IBSS)*, to fit our model.

In IBSS, we iteratively updating each "single effect" until it converges. The additional term $\mathbf{b}_0$ introduced to SuSiE-mixture can be viewed as another "single effect", thus we update $\mathbf{b}_0$ together with other "single effects".

Each updating step is an empirical Bayes method. When fitting each "single effect", we find the maximal likelihood estimator for σ_{0l}^2, and update the "single effect" with this MLE. For $\mathbf{b}_0$, we use a similar approach. But instead of maximizing the likelihood, which can be very complex, we maximize the *evidence lowerbound (ELBO)*. ELBO is a lowerbound of log-likelihood based on variational inference (Blei et al., 2017). See **Appendix B** for more details about maximizing the ELBO. The whole algorithm is showed as follows.

Algorithm 2 Fit SuSiE-mixture for unknown σ_b^2/σ^2

```
while not converge do
    $\bar{r} \leftarrow \mathbf{y} - \mathbf{X}\sum_{l=0}^{L}\bar{\mathbf{b}}_l$

    $\bar{r}_0 \leftarrow \bar{r} + \mathbf{X}\bar{\mathbf{b}}_0$
    $\sigma_b^2 \leftarrow \arg\max_{\sigma_b^2} ELBO_{ridge}(\bar{r}_0; \sigma_b^2, \sigma^2)$
    $(\bar{\mathbf{b}}_0, \overline{\mathbf{b}_0^2}) \leftarrow Ridge(\bar{r}_0, \mathbf{X}, \sigma^2, \sigma_b^2)$
    $\bar{r} \leftarrow \bar{r}_0 - \mathbf{X}\bar{\mathbf{b}}_0$

    for l in 1,..., L do
        $\bar{r}_l \leftarrow \bar{r} + \mathbf{X}\mathbf{b}_l$
        $\sigma_{0l}^2 \leftarrow \arg\max_{\sigma_0^2} l_{SER}(\bar{r}_l; \sigma_0^2, \sigma^2)$
        $(\alpha_l, \mu_{1l}, \sigma_{1l}) \leftarrow SER(\mathbf{X}, \bar{r}_l; \sigma^2, \sigma_{0l}^2)$
        $\bar{\mathbf{b}}_l \leftarrow \alpha_l \circ \mu_{1l}$
        $\overline{\mathbf{b}_l^2} \leftarrow \alpha_l \circ (\sigma_{1l}^2 + \mu_{1l}^2)$
        $\bar{r} \leftarrow \bar{r}_l - \mathbf{X}\bar{\mathbf{b}}_l$

    $\sigma^2 \leftarrow ERSS(\mathbf{y}, \bar{\mathbf{b}}, \overline{\mathbf{b}^2})/n.$
return $\sigma^2, \sigma_b^2, \boldsymbol{\sigma}_0^2, \boldsymbol{\alpha}, \boldsymbol{\sigma}_1, \boldsymbol{\mu}_1$
```

The inner loop and updating σ^2 are all steps in IBSS, and we can directly use them. The first few lines in the outer loop is simply a ridge regression. We can solve it analytically.

3.4 The choice of L

The result in Wang et al. (2018) shows, the key inferences of SuSiE are robust to overstating L. When L is larger than necessary, the method will be uncertain about where to place the extra effects, hence it will distribute them broadly among many variables. Therefore, this is going to have very little impact on the inference. For the SuSiE-mixture model, those reasons still hold. Thus in practice, we can appropriately over-estimate the value of L, and do not have to worry about that.

CHAPTER 4

SIMULATION

In the simulation, we set $n = 400, p = 1000, \sigma = 1, \sigma_b = 0.1, \sigma_{0l} = 5$, only 3 out of the 1000 variables are true factors. We run **Algorithm 1** with a large range of candidates of (σ_b/σ).

When we set the value of (σ_b/σ) in **Algorithm 1** to be exactly the true value 0.1, the following Figure 4.1 shows the result.

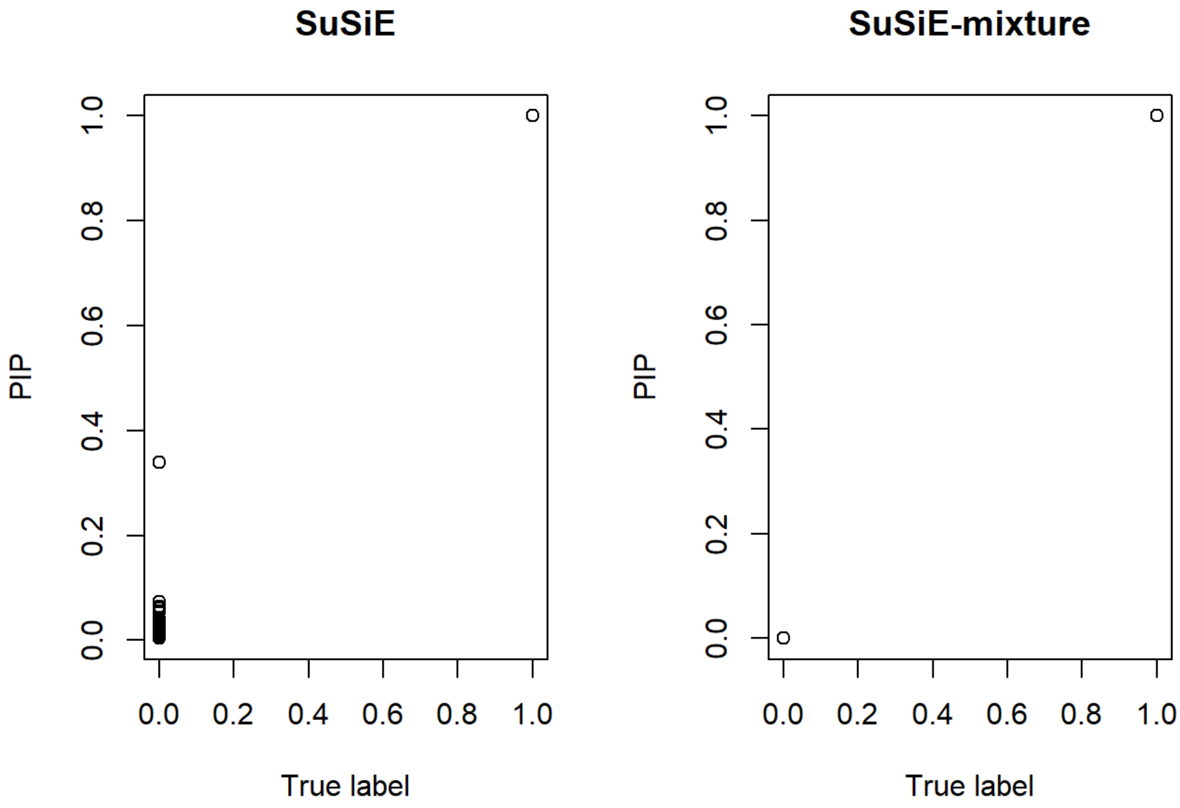

Figure 4.1: Compare the PIP of SuSiE and SuSiE-mixture. *True label=0* denotes the irrelevant variables, and *True label=1* denotes the true factors.

We can see that SuSiE has a lot of false discoveries: the PIP for those irrelevant variables are non-zero. And SuSiE-mixture shows a perfect selection, the PIP for irrelevant variables are close to 0, and the PIP for true factors are close to 1.

We calculate the KL-divergence between the true label and the predicted PIP. They both are equivalent to a Bernoulli variable. We can easily get the KL-divergence to be $-\log(\mathbb{P}\{\text{Correct prediction}\})$. And we sum up the KL-divergence for all variables to measure how good is the estimated PIP. A smaller KL-divergence indicates a better prediction.

Then we choose a wide range of (σ_b/σ) in **Algorithm 1**, and compare the KL-divergence with the SuSiE model. The result is showed in the following Figure 4.2.

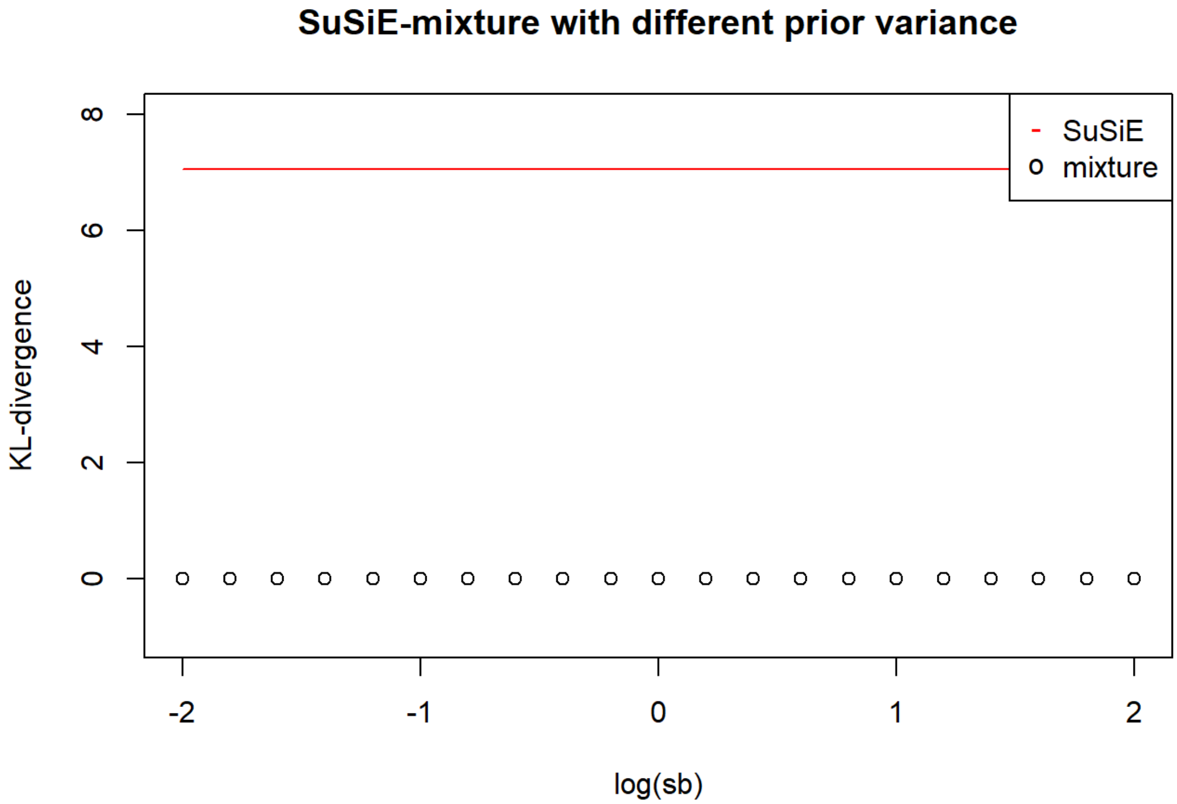

Figure 4.2: SuSiE-mixture with different prior variance

The result shows that the SuSiE-mixture model gets the perfect prediction within a wide range of choices of σ_b. The KL-divergence to the true label is almost constantly zero, much lower than the SuSiE model, indicating a better performance.

CHAPTER 5

DISCUSSION

We extend the spike-and-slab prior in the SuSiE model to mixture-Gaussian prior, and get more flexibility. We also show that the SuSie-mixture model has better performance in variable selection.

However, this method suffers from the following limitations. Normally the value of (σ_b^2/σ^2) is unknown to us, and we need to use **Algorithm 2** for fitting the SuSiE-mixture model. However, when the data has a large scale, it can be really slow. The ridge regression step in each iteration has $O(p^3)$ time complexity, but in real data set we often have $p > 10^4$, and the computation becomes a big trouble. Moreover, in the simulation study we find that SuSiE-mixture has a much worse convergence than SuSiE, taking much more iterations and time. Further studies to accelerate the fitting will be helpful.

REFERENCES

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Desboulets, L. (2018). A review on variable selection in regression analysis. *Econometrics*, 6(4):45.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573.

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS genetics*, 4(10):e1000214.

Wang, G., Sarkar, A. K., Carbonetto, P., and Stephens, M. (2018). A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, page 501114.

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264.

APPENDICES

A. Derivation of Algorithm 1

Assume the ratio of variance (σ_b^2/σ^2) is known. Let $\sum_{l=1}^{L}\mathbf{b}_l = \mathbf{b}$. Since

$$\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{X}\sum_{l=1}^{L}\mathbf{b}_l + \mathbf{e} = \mathbf{X}\mathbf{b}_0 + \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{5.1}$$

where

$$\mathbf{X}\mathbf{b}_0 \sim N(0, \sigma_b^2\mathbf{X}\mathbf{X}^T), \mathbf{e} \sim N(0, \sigma^2 I_n). \tag{5.2}$$

Thus

$$\mathbf{y}|\mathbf{X}, \mathbf{b}, \sigma^2, \sigma_b^2 \sim N\left(\mathbf{X}\mathbf{b}, \sigma_b^2\mathbf{X}\mathbf{X}^T + \sigma^2 I_n\right). \tag{5.3}$$

For simplicity, let

$$S = \left(\frac{\sigma_b^2}{\sigma^2}\mathbf{X}\mathbf{X}^T + I_n\right). \tag{5.4}$$

It is easy to see that S is positive definite. Then we find the Cholesky decomposition of S. Let L be a lower triangular matrix s.t. $LL^T = S$, i.e. $(L^{-1})^T L^{-1} = S^{-1}$. Further more we let $\tilde{\mathbf{y}} = L^{-1}\mathbf{y}$, $\tilde{\mathbf{X}} = L^{-1}\mathbf{X}$.

Then

$$\tilde{\mathbf{y}}|\mathbf{X}, \mathbf{b}, \sigma^2, \sigma_b^2 \sim N(L^{-1}\mathbf{X}\mathbf{b}, \sigma^2 L^{-1}S(L^{-1})^T). \tag{5.5}$$

That is,

$$\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \mathbf{b}, \sigma^2, \sigma_b^2 \sim N(\tilde{\mathbf{X}}\mathbf{b}, \sigma^2 I). \tag{5.6}$$

This formula together with the prior assumption on $\mathbf{b}$, will be exactly the SuSiE model setting

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{b} + \mathbf{e}, \mathbf{e} \sim N(0, \sigma^2 I_n), \tag{5.7}$$

We can easily solve this with $susie(\tilde{\mathbf{X}}, \tilde{\mathbf{y}}, L)$.

B. Derivation of variational inference method

Update σ_b^2 by maximizing the evidence lowerbound (ELBO)

In the ridge regression step, the ELBO can be written as

$$
\begin{aligned}
&ELBO(q_0) \\
=&\mathbb{E}_{q_0} \log p(X, \bar{r}_0; \beta) - \mathbb{E}_{q_0} \log q_0(\beta) \\
=&\mathbb{E}_{q_0} \left[\log(N(\bar{r}_0; X\beta, \sigma^2 I) \cdot N(\beta; 0, \sigma_b^2 I)) - \log q_0(\beta) \right] \\
=&\mathbb{E}_{q_0} \left[\log \left(\frac{\exp\left\{ (\hat{\beta}_{ridge}^T X^T \bar{r}_0 - \|\bar{r}_0\|_2^2)/(2\sigma^2) \right\}}{\sigma^n \sigma_b^p} \cdot N(\beta; \hat{\beta}_{ridge}, \hat{\Sigma}) \right) - \log q_0(\beta) \right] + const \\
=&(\hat{\beta}_{ridge}^T X^T \bar{r}_0 - \|\bar{r}_0\|_2^2)/(2\sigma^2) - n \log \sigma - p \log \sigma_b - KL(q_0 || N(\hat{\beta}_{ridge}, \Sigma_{ridge})) + const \\
\geq&(\hat{\beta}_{ridge}^T X^T \bar{r}_0 - \|\bar{r}_0\|_2^2)/(2\sigma^2) - n \log \sigma - p \log \sigma_b + const := G
\end{aligned}
$$

The equality holds when $q_0(\beta)$ is exactly $N(\hat{\beta}_{ridge}, \Sigma_{ridge})$.

Let

$$\frac{\partial G}{\partial \sigma_b} = \frac{1}{2\sigma^2} (\bar{r}_0^T X S^{-1} (\frac{2\sigma^2}{\sigma_b^3}) S^{-1} X^T \bar{r}_0) - \frac{p}{\sigma_b} = 0. \tag{5.8}$$

Thus

$$\|(\sigma_b^2 X^T X + \sigma^2 I)^{-1} X^T \bar{r}_0\|_2^2 = p. \tag{5.9}$$

We can numerically solve this with built-in function *optim()* in R.