

---

# BAYESIAN DECOUPLING FOR SPARSE ESTIMATION

---

Aihua Li

Department of Statistical Science

Duke University

**Advisor: Surya Tokdar, Jason Xu**

November, 2021

## ABSTRACT

The spike and slab prior, a gold standard for Bayesian variable selection, has poor sparsity behaviors on high dimensional dependent data. In high dimensional problems, Bayesian's intrinsic protect against model complexity imposes a favor for parsimonious models, and may lead to an over-conservative selection. Furthermore, when there is data dependency, marginal inclusion probabilities are scattered across correlated predictors, which often fail to determine an effective median probability model. Motivated by the pitfalls of spike and slab priors, this paper establishes Bayesian decoupling as an approach to sparse estimation on high dimensional data. Bayesian decoupling recovers the nature of Bayesian variable selection as a decision making procedure, and introduces a sparsity-inducing loss function penalizing model complexity. The goal is to find a sparse solution with a tolerable level of loss in the predictive ability from the best model averaging prediction. Our simulations show that on high dimensional data, the contribution of predictors in data variation is decaying, and we can identify few representative predictors explaining almost all variation. Showing how the decoupled sparsification gives a well-interpretable variable selection result, we will then focus on the control over false discovery rates via Bayesian decoupling in our future work.

**Keywords** Bayesian decoupling · Bayesian model selection · Spike and slab prior · Bernoulli-Gaussian mixture · Bayesian decision theory · Sparsity · High dimensional analysis · Linear regression

## 1 Introduction

Spike and slab prior is a standard approach for Bayesian model selection (see [1, 2]). It is a Bernoulli-Gaussian mixture, where the spike component places a point mass at zero, and the slab component is a Gaussian distribution. The spike and slab prior provides an desired uncertainty quantification over the  $2^p$  model space, with the caveats being the data dimension and data dependency. This paper will focus on the pitfalls of the spike and slab prior, and discuss its conservative behavior in high dimensional dependent cases.

Bayesian decoupling (BD), introduced in [3], provides an alternative option for Bayesian model selection. The idea of Bayesian decoupling comes from the challenge on whether a prior distribution can simultaneously provide the desired uncertainty quantification and imply a sparse interpretation. Instead, it decouples the task of Bayesian model fitting and the task of model selection, where the model selection is explicitly recovered as a decision making procedure with a sparsity-inducing loss.

The main purpose of this paper is to establish Bayesian decoupling as an approach to sparse estimation. We particularly focus on the high dimensional data where the spike and slab priors fail to concentrate the posterior inclusion probabilities and thus fail to determine an effective median probability model. The goal of Bayesian decoupling is to sparsify the Bayesian model averaging estimates, and decide on a sparse solution without too much loss in predictive ability compared to the BMA prediction. Our simulation shows that the decoupling can select a representative set of predictors which captures a high proportion of the total data variation. It provides an intermediate option to median probability models and full models, and leads to a well-interpretable sparse solution.

The rest of the paper is outlined as follows. Section 2 introduces the spike and slab prior for Bayesian model selection. Section 3 specifies the Bayesian decoupling approach. Section 4 conducts simulation to discuss the caveats of the spike and slab priors, and demonstrates Bayesian decoupling selection. Section 5 concludes the findings.

## 2 Spike and slab prior for Bayesian model selection

### 2.1 Linear regression model

Suppose for  $n$  observations  $\mathbf{y} = [y_1, \dots, y_n]^T$ , there are  $p$ -dimensional predictors  $\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]^T$  ( $i = 1, \dots, n$ ). Write the  $n \times p$  design matrix as  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ . Assume the columns of  $\mathbf{X}$  are mean-centered, so that  $\mathbf{X}^T \mathbf{1} = 0$ . Let  $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]^T$  be the random error.

A standard Gaussian linear regression model is given by

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (1)$$

where  $\mu$  is the intercept,  $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]^T$  is the  $p$ -dimensional regression coefficient, and  $\sigma^2 > 0$  is the variance of the Gaussian random error. Define  $k^* = \#\{i : \beta_i \neq 0\} = \|\boldsymbol{\beta}\|_0$  as the true model size.

## 2.2 Spike and slab prior

Spike and slab priors realizes the variable selection purpose by introducing a Bernoulli latent variable indicating the predictor's inclusion, and define hierarchically the Gaussian priors for coefficient  $\boldsymbol{\beta}$ . Therefore, spike and slab priors are essentially Bernoulli mixture of Gaussians. The slab component is typically a Gaussian distribution, while for the spike component, there are two general choices: *Dirac spikes*, which assign a point mass at zero, and *absolutely continuous spikes*, which places an another Guassian distribution centered at zero but much flatter than the slab component (see [4] for a complete discussion on the two types of spike and slab priors). In this paper, we focus on the Dirac spikes.

Formally, write the inclusion indicator as  $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_p]^p \in \{0, 1\}^p$ . For  $j = 1, \dots, p$ ,  $\gamma_j = 1$  indicates that  $\beta_j$  is allocated to the slab component and thus the  $j$ th predictor is selected in the model; otherwise  $\gamma_j = 0$ . For simplicity, we call the model determined by inclusion indicator  $\boldsymbol{\gamma}$  as "model  $\boldsymbol{\gamma}$ ".

Given the inclusion indicator  $\boldsymbol{\gamma}$ , the spike component is a point mass at zero, written as  $p(\beta_j = 0 | \gamma_j = 0) = \delta_0(\beta_j)$ .

For the slab component, assume a conjugate Gaussian prior with mean of  $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$  and covariance matrix  $\sigma^2 \boldsymbol{\Sigma}_{0,\boldsymbol{\gamma}}$ ,

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma}, \sigma^2 \sim N(\boldsymbol{\beta}_{0,\boldsymbol{\gamma}}, \sigma^2 \boldsymbol{\Sigma}_{0,\boldsymbol{\gamma}}).$$

In this paper, we consider  $\{\boldsymbol{\beta}_{0,\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{0,\boldsymbol{\gamma}}\} = \left\{ \mathbf{0}, n(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1} \right\}$ , where  $\mathbf{X}_{\boldsymbol{\gamma}}$  represents the design matrix consisting those columns of  $\mathbf{X}$  corresponding to  $\gamma_j = 1$ .

For the inclusion indicator  $\gamma_j$ 's, assume independent Bernoulli priors,

$$\gamma_j | \eta \stackrel{iid}{\sim} Bernoulli(\eta), \quad j = 1, \dots, p,$$

where  $\eta$  is usually called prior inclusion probability.  $\eta$  controls the overall sparsity level. A fixed  $\eta$  tends to result in a selection sensitive to the choice of the fixed number. Therefore, typically we define a hyper-prior for  $\eta$  as

$$\eta \sim beta(a, b), \quad (2)$$

and the default choice of  $a$  and  $b$  is the uniform prior  $a = b = 1$ .

Finally, a default uninformative prior for  $\{\mu, \sigma^2\}$  is

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

### 2.3 Bayesian model selection

After obtaining the posterior distribution from the spike and slab prior, typically there are three considerations in Bayesian model selection: 1) posterior probabilities of models, given by  $p(\gamma|\mathbf{y})$  for  $\gamma \in \{0, 1\}^p$ ; 2) posterior inclusion probabilities (also called marginal inclusion probabilities) of predictors, given by  $p(\gamma_j = 1|\mathbf{y})$  for  $j = 1, \dots, p$ ; and 3) Bayes factor, defined as  $BF(\gamma_1, \gamma_0) = \frac{p(\mathbf{y}|\gamma_1)}{p(\mathbf{y}|\gamma_0)}$ .

There are two common model selection criteria. One is to target at the highest posterior probabilities, and this will lead to the highest probability model (HPM). The other is to include all variables with posterior inclusion probability  $> 0.5$ , and then arrive at the median probability model (MPM). Sometimes people may also refer to the Bayes factor, especially to serve a hypothesis testing goal to compare two models. Bayes factor quantifies Bayesian's data evidence towards a model against the reference baseline model. Occasionally HPM and MPM are the same, while more often these two criteria will result in totally different selection. MPM are generally considered as having a better predictive ability than the HPM (see [5]).

## 3 Bayesian decoupling

### 3.1 Definition of Bayesian decoupling

Bayesian decoupling recovers the nature that Bayesian variable selection is a Bayesian decision making process. In the context of Bayesian decision theory, sparse estimation is an optimal action under a loss function minimizing the prediction error while penalizing the model complexity.

Formally, consider the prediction problem for new  $n$  observations  $\tilde{\mathbf{y}}$  given a new data matrix  $\tilde{\mathbf{X}}$ . Note that the new data matrix are pre-specified and can be different from the training data matrix  $\mathbf{X}$ . Nonetheless, in the rest of the paper we take  $\tilde{\mathbf{X}} = \mathbf{X}$  for notation simplicity.

Let  $\mathbf{b}$  be a coefficient estimate for the true regression coefficient  $\beta$ . To obtain a sparse solution, introduce a penalized squared prediction error loss given by

$$L(\beta, \mathbf{b}) = n^{-1} \|X\beta - X\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_c, \quad (3)$$

where  $\|\mathbf{b}\|_c$  controls the penalty for model complexity with tuning parameter  $\lambda \geq 0$ . (3) is termed as *decoupled shrinkage and selection (DSS)* loss function in [3].

By Bayesian decision theory, the Bayes estimate is given by the optimal solution which minimizes the posterior expectation of the DSS loss, written as

$$\mathbf{b}_\lambda^* = \arg \min_{\mathbf{b}} E_\beta [L(\beta, \mathbf{b}) | \mathbf{y}, X] = \arg \min_{\mathbf{b}} \int L(\beta, \mathbf{b}) p(\beta | \mathbf{y}, X) d\beta. \quad (4)$$

As shown in appendix A.1, (4) is equivalent to

$$\mathbf{b}_\lambda^* = \arg \min_{\mathbf{b}} n^{-1} \|X\bar{\beta} - X\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_c, \quad (5)$$

where  $\bar{\beta} = E_\beta [\beta | \mathbf{y}]$  is the posterior mean of  $\beta$ . Through the penalty for model complexity, the Bayes estimate (5) will realize the sparsity assumption. We call (5) as *Bayesian decoupling solution*.

### 3.2 Choices of the spenalty

Given posterior mean  $\bar{\beta}$ , the minimization in (5) comes down to a penalized optimization problem.

As any optimization question,  $L_0$  penalty poses a non-convex optimization difficulty, to which the closest convex relaxation is  $L_1$  penalty. Therefore, [3] implements the  $L_1$  optimization. Furthermore, due to the over-shrinking problem of  $L_1$  optimization, [3] considers the local linear approximation which solves a surrogate optimization problem (see [6, 7]) given by

$$\mathbf{b}_\lambda^* = \arg \min_{\mathbf{b}} n^{-1} \|X\bar{\beta} - X\mathbf{b}\|_2^2 + \sum_{j=1}^p \frac{\lambda}{|\bar{\beta}_j|} |b_j|. \quad (6)$$

In this paper, we consider both  $L_0$  and  $L_1$  optimization. For  $L_1$  penalty, we borrow the idea of [3] and implement the local linear alternative defined in (6). For  $L_0$  penalty, we realize the non-convex optimization via *iterative hard thresholding* (IHT) algorithm described in [8].

Specifically, for a standard  $L_0$  optimization problem,

$$\min_{\mathbf{b}} f(\mathbf{b}) = \frac{1}{n} \|\mathbf{y} - X\mathbf{b}\|_2^2 \quad \text{subject to } \|\mathbf{b}\|_0 \leq \lambda,$$

where  $\lambda$  takes integer value denoting the constraint on the model size, the IHT algorithm is to iterate

$$\begin{aligned} \mathbf{b}^{s+1} &= P_{S_k} [\mathbf{b}^s - \mu_s \nabla f(\mathbf{b}^s)] \\ &= P_{S_k} \left[ \mathbf{b}^s + \mu_s \times \frac{2}{n} X^T (\mathbf{y} - X\mathbf{b}^s) \right], \end{aligned}$$

where  $\nabla f(\mathbf{b}^s) = -\frac{2}{n} X^T (\mathbf{y} - X\mathbf{b}^s)$  is the gradient,  $\mu_s$  is the step size, and  $P_{S_k}(\mathbf{b})$  denotes the projection of  $\mathbf{b}$  onto the sparsity set  $S_k = \{\mathbf{b} : \|\mathbf{b}\|_0 \leq k\}$  where at most  $k$  components of a vector are nonzero. The projection is achieved by setting all but the  $k$  largest components of  $\mathbf{b}$  in magnitude equal to 0. The iteration continues until the tolerance level  $\frac{\|\mathbf{b}^s - \mathbf{b}^{s-1}\|^2}{\|\mathbf{b}^s\|^2} < 10^{-8}$ .

At each iteration step, we consider the step size suggested in [8], given by

$$\mu_s = \frac{\|\mathbf{b}^s\|_2^2}{\|X\mathbf{b}^s\|_2^2}.$$

### 3.3 Evaluating the sparsification

In the decoupling solution (5), the tuning parameter  $\lambda$  controls the sparsity level. There are two extremes. On one extreme,  $\lambda = 0$  recovers the posterior mean estimate  $\mathbf{b}_0^* = \bar{\beta}$ . On the other extreme,  $\lambda = \infty$  implies a complete shrinkage towards zero, i.e.,  $\mathbf{b}_\infty^* = \mathbf{0}$ . Therefore, the Bayesian decoupling is a sparsification over the posterior mean  $\bar{\beta}$  towards  $\mathbf{0}$ .

Furthermore, as shown in appendix A.2, the posterior mean  $\bar{\beta}$  of the spike and slab prior leads to the Bayesian model averaging (BMA) prediction  $E[\tilde{\mathbf{y}}|\mathbf{y}] = X\bar{\beta}$ . The BMA thus provides the best prediction in terms of the squared error loss  $n^{-1}\|X\beta - X\mathbf{b}\|_2^2$ . Then, it shapes our key question: compared to the BMA prediction, how much predictive deterioration has Bayesian decoupling induce along the sparsification path? The ultimate goal of Bayesian decoupling for sparse estimation is to find a sparse solution with a tolerable level of loss in the predictive ability from the best model averaging prediction.

Introduced in [3], the *selection summary plots* provide a way of evaluation by visualizing the change in predictive ability for different choices of  $\lambda$ .

The first plot is the path of *variation-explained*, defined as

$$\rho_\lambda^2 = \frac{n^{-1} \|\mathbf{X}\boldsymbol{\beta}\|^2}{n^{-1} \|\mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 + n^{-1} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\mathbf{b}_\lambda^*\|^2}. \quad (7)$$

Compared to the "benchmark"

$$\rho^2 = \frac{n^{-1} \|\mathbf{X}\boldsymbol{\beta}\|^2}{n^{-1} \|\mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2},$$

$\rho_\lambda^2$  quantifies the decrease in the variation-explained due to the additional noise  $n^{-1} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\mathbf{b}_\lambda^*\|^2$ .

The second plot is the path of *excess error*, defined as

$$\psi_\lambda = \sqrt{n^{-1} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\mathbf{b}_\lambda^*\|^2 + \sigma^2} - \sigma. \quad (8)$$

The third plot is the suggested being the coefficient solution path, which plots the magnitude of  $\mathbf{b}_{\lambda,j}^*$ 's versus the model size  $\|\mathbf{b}_\lambda^*\|_0$ , the excess error  $\psi_\lambda$ , etc.

Note that both  $\rho_\lambda^2$  and  $\psi_\lambda$  are random variable, with the randomness coming from the posterior of  $\boldsymbol{\beta}$ . Then the 95% credible intervals of  $\rho_\lambda^2$  and  $\psi_\lambda$  quantify the uncertainty in the two metrics. A heuristic approach for determining a single model is to choose the smallest model whose 95% upper bound of  $\rho_\lambda$  achieves the benchmark level  $\rho^2$  (see [3]).

### 3.4 A decision theory viewpoint of Bayesian model selection

Bayesian model selection is implicitly a hypothesis testing problem, and selecting the most significant set of predictors is indeed finding the optimal action under a targeted loss.

In fact, the HPM is the optimal action under  $L(\boldsymbol{\beta}, \mathbf{b}) = 1 (\boldsymbol{\beta} = \mathbf{b}) = l_0(\boldsymbol{\beta}, \mathbf{b})$ , while the MPM is the optimal action under  $L(\boldsymbol{\beta}, \mathbf{b}) = \sum_{j=1}^p |\beta_j - b_j| = l_1(\boldsymbol{\beta}, \mathbf{b})$  (see [3]). We can note that in these two loss functions, there is intrinsically no sparsity-inducing mechanism. Both of them seeks the set of predictors with a minimum distance to the true coefficient  $\boldsymbol{\beta}$  whose randomness is quantified by the posterior. In other words, we rely on the spike and slab prior to provide an uncertainty quantification (for both the randomness in the coefficient and the randomness in the model choice), while simultaneously hoping the quantification implies a sparse interpretation.

This is doable in some ideal cases where the posterior distribution is distinctly concentrated around the true model. However, in more general situations where the posterior of the spike and slab priors fails to concentrate, the above two losses can fail to select a meaningful model. For example, the signal-to-noise ratio is an important consideration in

spike and slab selection (see [9]), and we typically require the signal to be clear enough to avoid a harsh shrinkage towards zero.

On the other hand, Bayesian decoupling explicitly applies the nature of Bayesian model selection as a decision making procedure, and imposes the sparsity assumption via a sparsity-inducing loss. The decoupling discards the reliance on the concentration of the posterior, and seeks for the most predictive small set of predictors from the model averaging gold standard.

In addition, since the task of Bayesian model fitting and the task of model selection are separated, decoupling relaxes the choice of the prior. Instead of being restricted to the selection prior such as the spike and slab prior to induce sparsity, we can consider other alternatives to provide the desired uncertainty quantification.

## 4 Simulation

### 4.1 Pitfalls of spike and slab priors

Our simulation is motivated by an application of spike and slab priors on Tecator data in [10], where the extremely highly correlated predictors severely split the posterior inclusion probabilities. In this case, almost none of the predictors can have an inclusion probability  $> 0.5$  to form an effective median probability model, despite the true association between the predictors and the response. Therefore, this section investigates into the performance of the spike and slab prior on high dimensional dependent data, with a particular focus on the distribution of the posterior inclusion probabilities (PIP) and the corresponding size of the median probability models (MPM).

We consider the following data generating model,

$$y_i = \mu + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \mathbf{x}_i \stackrel{iid}{\sim} N(\mathbf{0}, R), \quad \epsilon_i \stackrel{iid}{\sim} N(0, 1), \quad i = 1, \dots, n, \quad (9)$$

where the  $p$ -dimensional predictors are pairwise correlated, controlled by the covariance matrix

$$R = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{p \times p}.$$

The true coefficient vector is

$$\boldsymbol{\beta} = \left[ \underbrace{1 \cdots 1}_{\text{true model size } k^*} \quad \underbrace{0 \cdots 0}_{p-k^*} \right]^T \in \mathbb{R}^{p \times 1}.$$

In this section, we first consider sample size  $n = 100$ , dimension  $p = 20, 100, 200$ , true model size  $k^* = 5, 15$ , and correlation  $\rho = 0.2, 0.5, 0.7, 0.9$ . The spike and slab model fitting is implemented by `scaleBVS` package in R, which codes the weighted tempered Gibbs sampling algorithm proposed in [11]. We conduct 100 repeated simulations on each parameter setting, while in each simulation, there are 10000 MCMC samples with 1000 burn-in samples.

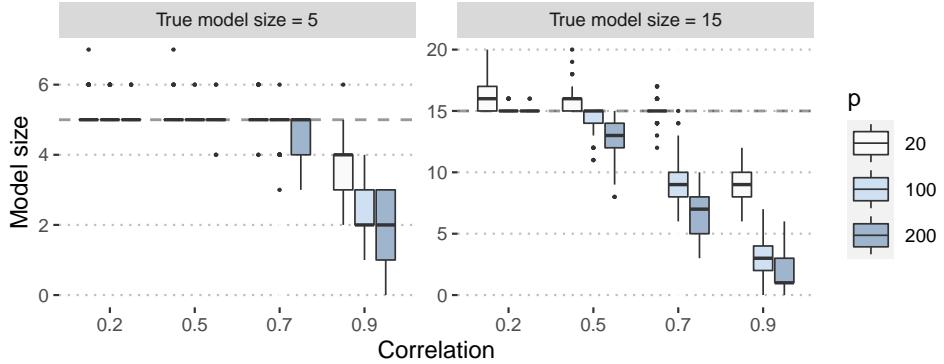


Figure 1: Model sizes in MPM with spike and slab prior in 100 repeated simulations, where  $n = 100$

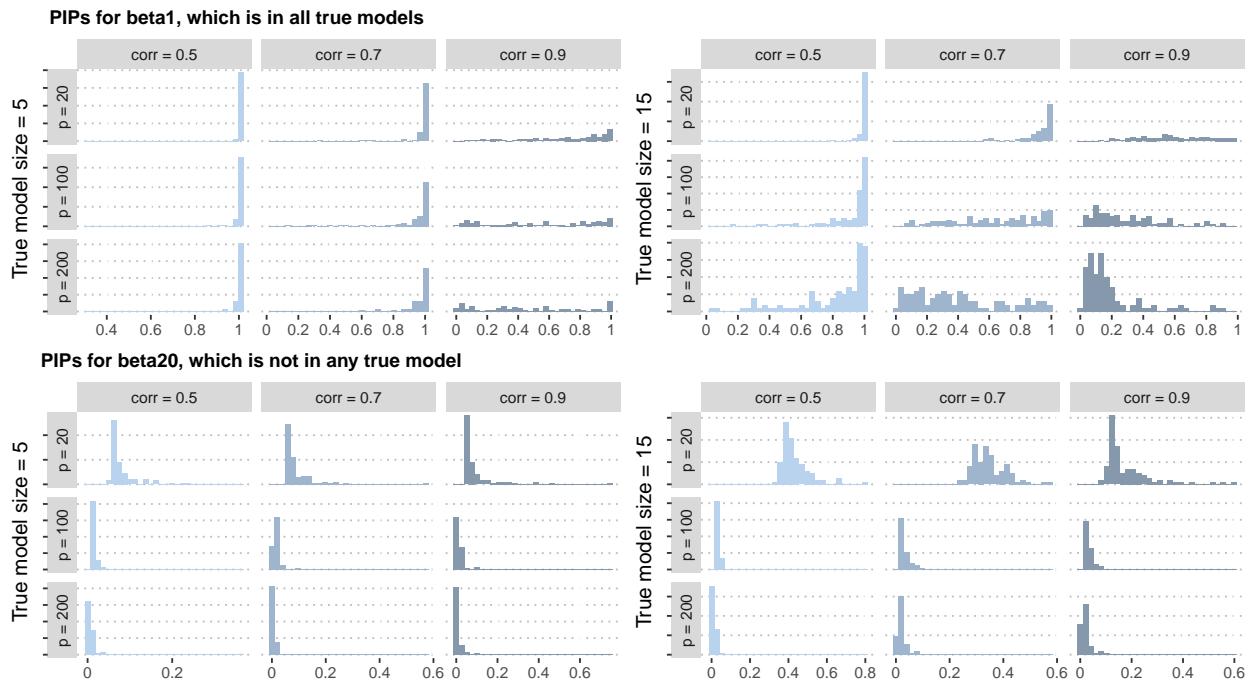


Figure 2: Histograms of posterior inclusion probabilities in the 100 repeated simulations for two predictors

Figure 1 shows the model sizes of median probability models (i.e., number of predictors with posterior inclusion probability  $> 0.5$ ) in different simulation settings. Figure 2 shows the histograms of the posterior inclusion probabilities in the 100 repeated simulations for two randomly chosen predictors, where one is a true signal in all settings and the other is a zero signal.

#### 4.1.1 Correlation

No matter the true model size and the dimension, an extremely correlated data will make the spike and slab prior fail. As shown in 2, for the true signal  $\beta_1$ , the distribution of PIPs in the repeated simulations is "flat" across  $[0, 1]$  when  $\rho = 0.9$ , which means that we generally have a half chance to select this true signal into the model and a half chance not.

In general, a higher correlation makes the spike and slab prior favor more over the smaller models. Besides, in the current simulations where the true model size is small compared to the dimension, this happens when there is at least a moderate level of correlation (e.g.,  $\rho = 0.5 - 0.7$ ), while a low level of correlation (e.g.,  $\rho = 0.2 - 0.5$ ) won't affect the spike and slab prior much. As we will see in the following discussions in figure 3, when the true model size is large, any increase in the correlation will lead to a dramatic increase in the shrinkage power towards smaller models.

#### 4.1.2 Dimension

A higher dimension leads to the preference of the spike and slab prior towards smaller models.

According to figure 1, in low dimensional case where  $p = 20$ , the model sizes in MPMs are roughly the same as the true model size, except for the extreme case where  $\rho = 0.9$ . Besides, notably, in the low dimensional case, the spike and slab prior may slightly overestimate the model size. As shown in figure 2, for  $\beta_{20}$  which is indeed a zero signal, the PIPs under  $p = 20$  are not concentrated at 0 but move towards the right. Especially when the true model size is large, there is a higher chance that spike and slab prior will select a model larger the true model.

The most notable change happens when the dimension increases from  $p \ll n$  to  $p \approx n$ , given that there is a moderately large correlation and the true model size is not too small. As shown in figure 1 and 2, when  $\rho = 0.7, 0.9$  and true model size = 15, the model sizes of MPMs dramatically drop down when  $p$  grows from  $20 < n$  to  $100 = n$ , and the distribution of PIPs for  $\beta_1$  flattens out from the spike at 1.

However, starting from  $p \approx n$ , the further increase in dimension doesn't bring a same level of further shrinkage in the model size. As shown in figure 1, for  $\rho = 0.7, 0.9$ , the decrease in model sizes when  $p$  goes from 100 to 200 is not as much as that when  $p$  goes from 20 to 100.

#### 4.1.3 True model size

The true model size greatly affects the performance of spike and slab prior.

When the true model size is small, unless there is an extremely large correlation among the data, the varying dimension and correlation won't much affect the performance of spike and slab prior. As shown in figure 1 and 2, the true signals can always stand out and the selected model sizes in MPM are basically the true model size.

On the contrary, the selection becomes vague for the spike and slab prior when the true model size is large, and this is where the increasing correlation and dimension can have a clear impact on the shrinkage power.

Motivated by this, we take a further look at the impacts of the true model size. Figure 3 shows the model sizes in MPMs in 100 repeated simulations. Here  $n = 50$ , and we consider  $p = n$  and  $p = 2n$ . Also, we consider a wide range of correlation from 0 to 0.7.

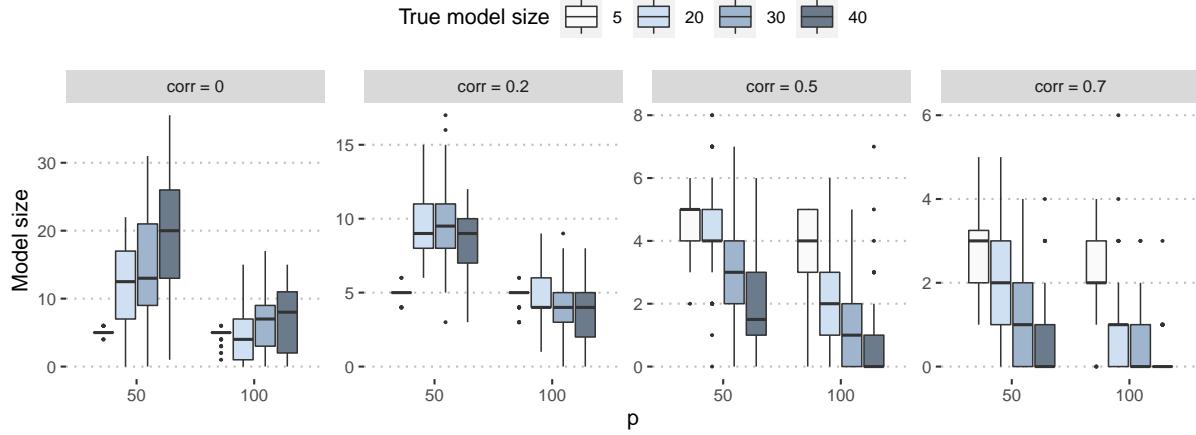


Figure 3: Model sizes in MPM with spike and slab prior in 100 repeated simulations, where  $n = 50$

As mentioned above, the increasing dimension and the increasing correlation impose a preference towards the more parsimonious models. In addition to this, figure 3) shows the shrinkage power of spike and slab prior in large models, even in independent designs. More specifically, in cases where  $\rho = 0$ , the selected model sizes won't achieve the true model sizes. When the correlation increases, the selected models are much smaller than the true models. The extreme

situation is where the correlation is large and the dimension is high, and we can often see spike and slab prior select nearly null models.

#### 4.1.4 An interpretation on spike and slab priors

Bayesian intrinsic penalty for model complexity lies in the computation of marginal likelihood. While the (conditional) likelihood favors complicated models which maximize the likelihood, the marginal likelihood reweights the likelihood according to the priors. Then, Bayesian model selection, which is based on marginal likelihoods, provides an implicit protection against huge models, as averaging over a huge parameter space makes the marginal likelihood less concentrated around its highest value.

In our simulation, we have seen that the increasing dimension leads to a higher preference towards smaller models, although the true model size is indeed the same. The enlarged data brings a bunch of additional potential predictors into consideration, which pushes the maximum conditional likelihood further towards complicated models, while indeed the improvement in the conditional likelihood due to these nuisance predictors can be tiny. Meanwhile, the additional integration over the additional dimensions provides more smoothing on the integrand, resulting in a flatter marginal likelihood. In this case, smaller models win in the trade-off between the likelihood and the model complexity.

Besides the view of Bayesian's intrinsic penalty for model complexity via marginal likelihoods, another interpretation of the favor towards parsimonious models comes from the implicit multiple testing problem in Bayesian model selection. When testing whether each of the predictors has a non-zero association with the response, the overall inclusion probability provides a control over the overall error rates. Then, targeting at lower overall error rates, the posterior of the overall inclusion probability in high dimensional cases will be concentrated around an extremely small value close to 0. Therefore, as we see in the simulations, an increasing dimension leads to a more conservative selection, making it hard for any of the marginal posterior inclusion probabilities to stand out.

In addition to the impacts of dimension, data dependency also reduces the selected model size. Strong data dependency makes it difficult to differentiate among the contributions of correlated predictors on the response variation. Unlike the orthogonal case where the association between each predictor and the response is distinct, the correlated predictors in part share a common association. Especially when the correlation level is high, the joint inclusion probability of the correlated predictors as a group is shared by each component, so that none of the individual predictors in the group will have a high inclusion probability. Although this can be a beneficial property in some cases where a sparse model is desired for interpretation purposes, we should be cautious when 1) the dependency is strong, and 2) the correlated

group is large, as in this case, with the inclusion probability split over the large group, the median probability model can be the null model.

The above also answers why the spike and slab prior can always select the true signals when the true model size is small, even if there is data dependency. As discussed in the simulation section, except for the extreme case where the correlation is nearly 1, a tiny group of true signals collecting all explaining power are always able to stand out. On the contrary, when a large number of predictors all have non-zero associations with the response, then, together with the issues of high dimension and data dependency, the spike and slab prior tends to underestimate the individual involvement. Confused by too many vague individual contributions, the spike and slab prior prefers the safer option and sticks to parsimonious models.

## 4.2 Bayesian decoupling for sparse estimation

Motivated by the pitfalls of spike and slab priors, this section shows how Bayesian decoupling determines a sparse solution on high dimensional data.

Consider the same data generating model defined in (9). In this section, we take sample size  $n = 50$ , dimension  $p = 50$ , true model size  $k^* = 5, 20, 40$ , and correlation  $\rho = 0, 0.2, 0.5, 0.7$ . In any MCMC sampling, we draw 10000 samples with 1000 burn-in samples. We consider both  $L_0$  and  $L_1$  penalties, where the  $L_1$  optimization is realized by `lars` package in R, and for the  $L_0$  optimization, we implement the IHT algorithm described in section 3.2.

Appendix B.2 shows the selection summary plots of single simulations under different settings, and additionally plots the decoupling solutions versus the posterior means and the posterior inclusion probabilities for comparison. Figure 4 collects in one place the paths of variation-explained  $\rho_\lambda^2$  and excess error  $\psi_\lambda$  by  $L_0$  decoupling in different simulation settings. Figure 5 shows the values of different metrics in 50 repeated simulations, with  $n = p = 50$  and true model size = 30.

### 4.2.1 Solution paths of $L_0$ versus $L_1$

In general,  $L_1$  selects predictors according to their posterior means. That is, predictors with larger posterior means will enter the decoupling model at first, and in the final selected model, the coefficients estimates are generally proportional to the posterior means. In addition, in the independent cases or the low correlation cases, due to  $L_1$ 's harsh penalty through large  $\lambda$ , the coefficients are generally underestimated at the beginning of the decoupling path, and gradually "grow" to the posterior mean levels. For example, see the two independent cases in figure 11 and 15.

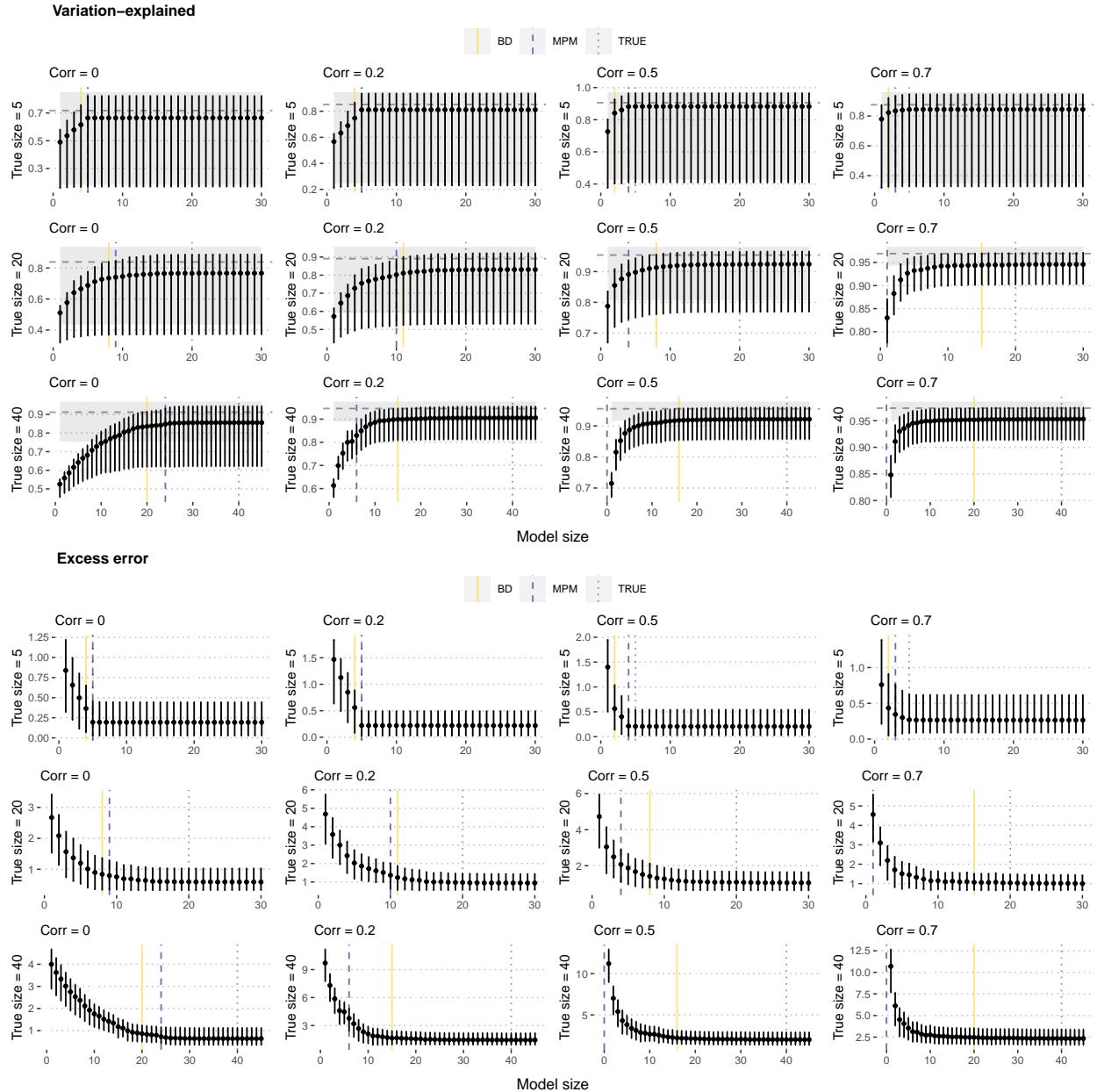


Figure 4: The paths of variation-explained  $\rho_\lambda^2$  and excess error  $\psi_\lambda$  by  $L_0$  decoupling in different simulations with  $n = 50, p = 50$

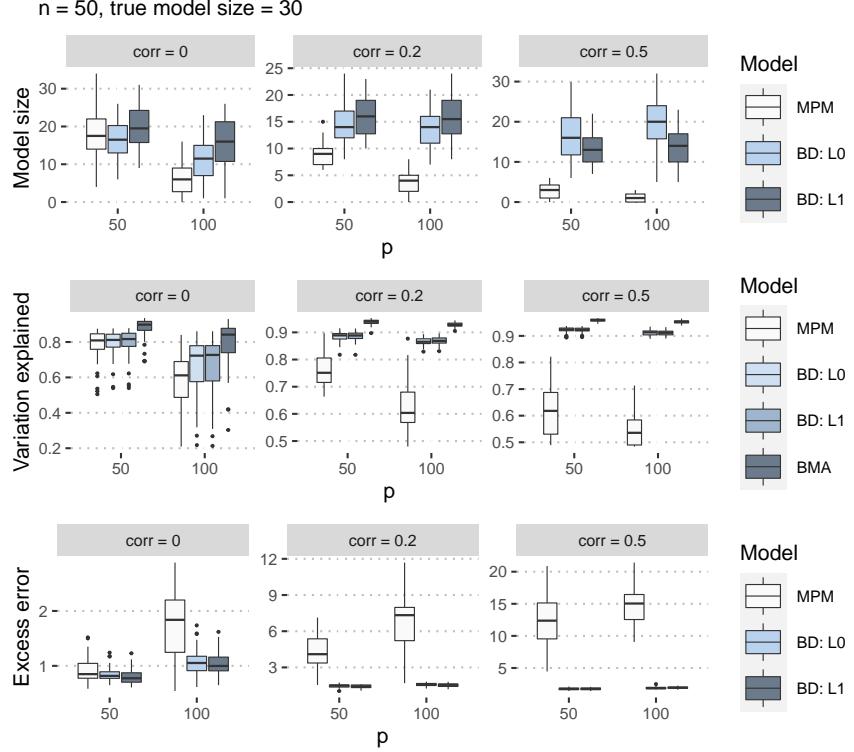


Figure 5: Different decoupling metrics in 50 repeated simulations with  $n = p = 50$  and true model size = 30

On the contrary,  $L_0$  doesn't necessarily select the predictors according to the posterior inclusion probabilities or the posterior means. As shown in 8, even in the simplest design where there are only 5 true signals and their PIPs are perfectly 1, the  $L_0$  optimization has a chance to first select predictors with low PIPs. In general, this is more likely to happen when the data are dependent, as in figure 12 and 13 where the correlation grows, there are more predictors with low posterior means selected along the solution path. Also, in the dependent case, the coefficient estimates in the final selected model may not be proportional to the posterior means. In addition to this,  $L_0$  allows more overestimation early on the solution path because it doesn't penalize the estimates magnitudes.

Besides, there can be different predictors in and out along the solution path of  $L_0$ , while  $L_1$  almost always sticks to the predictors already selected, and simply change their magnitudes along the solution path. For this reason, the solution path of  $L_0$  looks a little messy.

In conclusion,  $L_0$  more "randomly" decides its predictors, especially when there is some correlation among the data. One concern is that it may not be a problem of  $L_0$  itself but a problem of the IHT algorithm. In fact, it is known that when the signals are weak, it is hard for IHT to find the global optimum. In the BD application, the posterior means are all pretty small especially when the spike and slab prior fails to concentrate the PIPs. Thus, we cannot rule out the

possibility that the "arbitrary selection" behavior of  $L_0$  decoupling is indeed because the IHT fails to find the optimal solution.

#### 4.2.2 Decaying contribution to data variation

As shown in figure 4, in small models (i.e., model with 5 true signals), the variation explained  $\rho_\lambda^2$  and the excess error  $\psi_\lambda$  are improved "linearly"; that is, the first few predictors have equal contributions to these two measurements. However, when the true model size grows, the paths of both  $\rho_\lambda^2$  and  $\psi_\lambda$  show a "curve" improvement; that is, in general, the improvement in  $\rho_\lambda^2$  and  $\psi_\lambda$  is "marginally decreasing" as more predictors enter the model. Besides, as the correlation increases, the turning point comes earlier, which means that a small number of predictors can capture a large proportion of the variation.

It is a nice result, in the sense that when there are truly many competing signals associated with the response so that the full model with BMA estimates is too large but the MPM may completely reduce to a null model, BD provides an intermediate option. In this case, the first few number of predictors can have a most significant contribution to the data variation, while including the rest of the predictors only brings a decaying improvement in explaining the variation.

#### 4.2.3 Uncertainty quantification of the decoupling metrics

The increasing true model size and the increasing correlation make spike and slab prior fails to shape a concentrated posterior over the model space. However, it turns out that the uncertainty in the variation explained  $\rho_\lambda^2$  and excess error  $\psi_\lambda$  reduces when the true model size and the correlation level grow.

As shown in figure 4, an increasing true model size and an increasing correlation will bring in larger excess errors ( $\psi_\lambda = \sqrt{n^{-1}||\mathbf{X}\beta - \mathbf{X}b_\lambda^*||^2 + \sigma^2} - \sigma$ ) when there is a strong sparsification. However, the counterintuitive thing is that the variation explained ( $\rho_\lambda^2 = \frac{n^{-1}||\mathbf{X}\beta||^2}{n^{-1}||\mathbf{X}\beta||^2 + \sigma^2 + n^{-1}||\mathbf{X}\beta - \mathbf{X}b_\lambda^*||^2}$ ) improves at the same time – its magnitude increases closer to 1 and its uncertainty decreases.

A possible reason is that when the PIPs fail to concentrate and all predictors have competing posterior means, a non-sparse coefficient  $\beta$  results in a large total variation quantified by  $n^{-1}||\mathbf{X}\beta||^2$ . That is why  $\rho^2$  improves in these "bad cases" for spike and slab prior. On the other hand, an increasing  $\psi_\lambda$  implies larger  $n^{-1}||\mathbf{X}\beta - \mathbf{X}b_\lambda^*||^2$ , which means that we indeed suffer from a large excess error in prediction when doing sparsification over this huge true model.

#### 4.2.4 Findings in repeated simulations

See figure 5. As for the model size, when the MPM fails in cases with 1) large true model sizes, 2) high dimension, and/or 3) high data dependency, BD can recover a non-null model.  $L_0$  or  $L_1$  won't definitely choose a smaller or larger model than each other. It turns out that when the correlation is weak,  $L_0$  tends to select a smaller model, while when the correlation is strong,  $L_0$  will result in a larger model than  $L_1$ .

The results of variation explained and excess error are not surprising, as the BD models are determined by a fixed tolerance in the variation-explained lost during sparsification. In other words, the variation explained of BD models will always be 95% of the BMA models. Also, MPMs will have the lowest variation explained and higher excess error, which is not surprising, as we consider simulation settings where the spike and slab prior may fail.

## 5 Conclusions

The spike and slab prior, a gold standard for Bayesian variable selection, leads to an over conservative selection on high dimensional dependent data. Due to Bayesian's intrinsic protect against model complexity, the spike and slab prior unduly underestimates the inclusion probabilities in high dimensional cases. Especially when there are further the data dependencies, the posterior inclusion probabilities are severely splitted across the correlated group, and thus fail to determine an effective median probability model.

Motivated by the pitfalls of spike and slab priors, this paper establishes Bayesian decoupling as an alternative approach to sparse estimation. Bayesian decoupling recovers the nature of Bayesian variable selection as a decision making procedure, and introduces a sparsity-inducing loss penalizing model complexity. It provides a sparsification over the Bayesian model averaging estimates, and is targeted at a sparse solution with a tolerable level of loss in the predictive ability from the best model averaging prediction.

Our simulation on high dimensional data shows a decaying contribution of predictors to explaining the variation in the BMA prediction. The first few number of predictors can have a most significant contribution, while including the rest of the predictors only brings a reduced improvement. In addition, the presence of data dependency can shrink the selected model size, as fewer representative predictors can fully capture the data variation. In conclusion, Bayesian decoupling provides an intermediate option to MPM and BMA, and leads to a well-interpretable sparse solution.

## References

- [1] T. J. Mitchell and J. J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, December 1988.
- [2] Edward I. George and Robert E. McCulloch. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889, September 1993.
- [3] P. Richard Hahn and Carlos M. Carvalho. Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective. *Journal of the American Statistical Association*, 110(509):435–448, January 2015.
- [4] Gertraud Malsiner-Walli and Helga Wagner. Comparing Spike and Slab Priors for Bayesian Variable Selection. *arXiv:1812.07259 [stat]*, December 2018. arXiv: 1812.07259.
- [5] Maria Maddalena Barbieri and James O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3), June 2004.
- [6] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4), August 2008.
- [7] Jinchi Lv and Yingying Fan. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A), December 2009.
- [8] Kevin L. Keys, Gary K. Chen, and Kenneth Lange. Iterative hard thresholding for model selection in genome-wide association studies. *Genetic Epidemiology*, 41(8):756–768, December 2017.
- [9] Nicholas G. Polson and Lei Sun. Bayesian l0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, May 2019.
- [10] J E Griffin, K G Łatuszyński, and M F J Steel. In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large  $p$ . *Biometrika*, 108(1):53–69, March 2021.
- [11] Giacomo Zanella and Gareth Roberts. Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):489–517, July 2019.

## A Appendix: Derivation

### A.1 Derivation of the Bayesian decoupling solution

Suppose we are given a pre-specified (new) data matrix  $\mathbf{X}$ . To obtain the Bayes estimate under the DSS loss defined in (3), first take expectation with respect to the posterior  $p(\boldsymbol{\beta}|\mathbf{y})$ , then we have

$$\begin{aligned} E_{\boldsymbol{\beta}} [L(\boldsymbol{\beta}, \mathbf{b})|\mathbf{y}] &= E_{\boldsymbol{\beta}} [n^{-1}\|X\boldsymbol{\beta} - X\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_c|\mathbf{y}, \mathbf{X}] \\ &= n^{-1}E_{\boldsymbol{\beta}} [\|X\boldsymbol{\beta} - X\mathbf{b}\|_2^2|\mathbf{y}] + \lambda\|\mathbf{b}\|_c. \end{aligned} \quad (10)$$

Write the posterior mean of  $\boldsymbol{\beta}$  as  $\bar{\boldsymbol{\beta}} = E_{\boldsymbol{\beta}} [\boldsymbol{\beta}|\mathbf{y}]$ . Note that

$$\begin{aligned} E_{\boldsymbol{\beta}} [\|X\boldsymbol{\beta} - X\mathbf{b}\|_2^2|\mathbf{y}] &= E_{\boldsymbol{\beta}} [\|X\boldsymbol{\beta} - X\bar{\boldsymbol{\beta}} + X\bar{\boldsymbol{\beta}} - X\mathbf{b}\|_2^2|\mathbf{y}] \\ &= E_{\boldsymbol{\beta}} [\|X\boldsymbol{\beta} - X\bar{\boldsymbol{\beta}}\|_2^2 + 2(X\boldsymbol{\beta} - X\bar{\boldsymbol{\beta}})^T(X\bar{\boldsymbol{\beta}} - X\mathbf{b}) + \|X\bar{\boldsymbol{\beta}} - X\mathbf{b}\|_2^2|\mathbf{y}] \\ &= E_{\boldsymbol{\beta}} [\|X\boldsymbol{\beta} - X\bar{\boldsymbol{\beta}}\|_2^2|\mathbf{y}] + \|X\bar{\boldsymbol{\beta}} - X\mathbf{b}\|_2^2. \end{aligned}$$

Therefore, the expected loss in 10 is

$$E_{\boldsymbol{\beta}} [L(\boldsymbol{\beta}, \mathbf{b})|\mathbf{y}] = n^{-1}E_{\boldsymbol{\beta}} [\|X\boldsymbol{\beta} - X\bar{\boldsymbol{\beta}}\|_2^2|\mathbf{y}] + n^{-1}\|X\bar{\boldsymbol{\beta}} - X\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_c.$$

Drop the first term which is constant in the action  $\mathbf{b}$ , then the Bayes estimate is

$$\mathbf{b}_\lambda^* = \arg \min_{\mathbf{b}} E_{\boldsymbol{\beta}} [L(\boldsymbol{\beta}, \mathbf{b})|\mathbf{y}, \mathbf{X}] = \arg \min_{\mathbf{b}} n^{-1}\|X\bar{\boldsymbol{\beta}} - X\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_c.$$

### A.2 A Bayesian model averaging viewpoint of the spike and slab posterior mean

Given a pre-specified (new) data matrix  $\mathbf{X}$ , the model averaging prediction is

$$\begin{aligned} E[\tilde{\mathbf{y}}|\mathbf{y}] &= \sum_{\boldsymbol{\gamma} \in \{0,1\}^p} E[\tilde{\mathbf{y}}|\mathbf{y}, \boldsymbol{\gamma}] p(\boldsymbol{\gamma}|\mathbf{y}) \\ &= \sum_{\boldsymbol{\gamma} \in \{0,1\}^p} E[\mathbf{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}|\mathbf{y}, \boldsymbol{\gamma}] p(\boldsymbol{\gamma}|\mathbf{y}) \\ &= \mathbf{X} \sum_{\boldsymbol{\gamma} \in \{0,1\}^p} E[\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\gamma}] p(\boldsymbol{\gamma}|\mathbf{y}). \end{aligned}$$

Note that the model averaging estimate is

$$\begin{aligned}
\sum_{\gamma \in \{0,1\}^p} E[\beta | \mathbf{y}, \gamma] p(\gamma | \mathbf{y}) &= \sum_{\gamma \in \{0,1\}^p} \left( \int \beta \cdot p(\beta | \mathbf{y}, \gamma) d\beta \right) p(\gamma | \mathbf{y}) \\
&= \int \beta \cdot \left( \sum_{\gamma \in \{0,1\}^p} p(\beta | \mathbf{y}, \gamma) p(\gamma | \mathbf{y}) \right) d\beta \\
&= \int \beta \cdot p(\beta | \mathbf{y}) d\beta \\
&= E[\beta | \mathbf{y}] \\
&= \bar{\beta}.
\end{aligned}$$

It implies that the posterior mean  $\bar{\beta}$  from the spike and slab prior is the Bayesian model averaging estimate, and the corresponding prediction  $\mathbf{X}\bar{\beta}$  is the Bayesian model averaging prediction.

## B Appendix: Additional plots

### B.1 A visualization of the posterior inclusion probabilities over all predictors

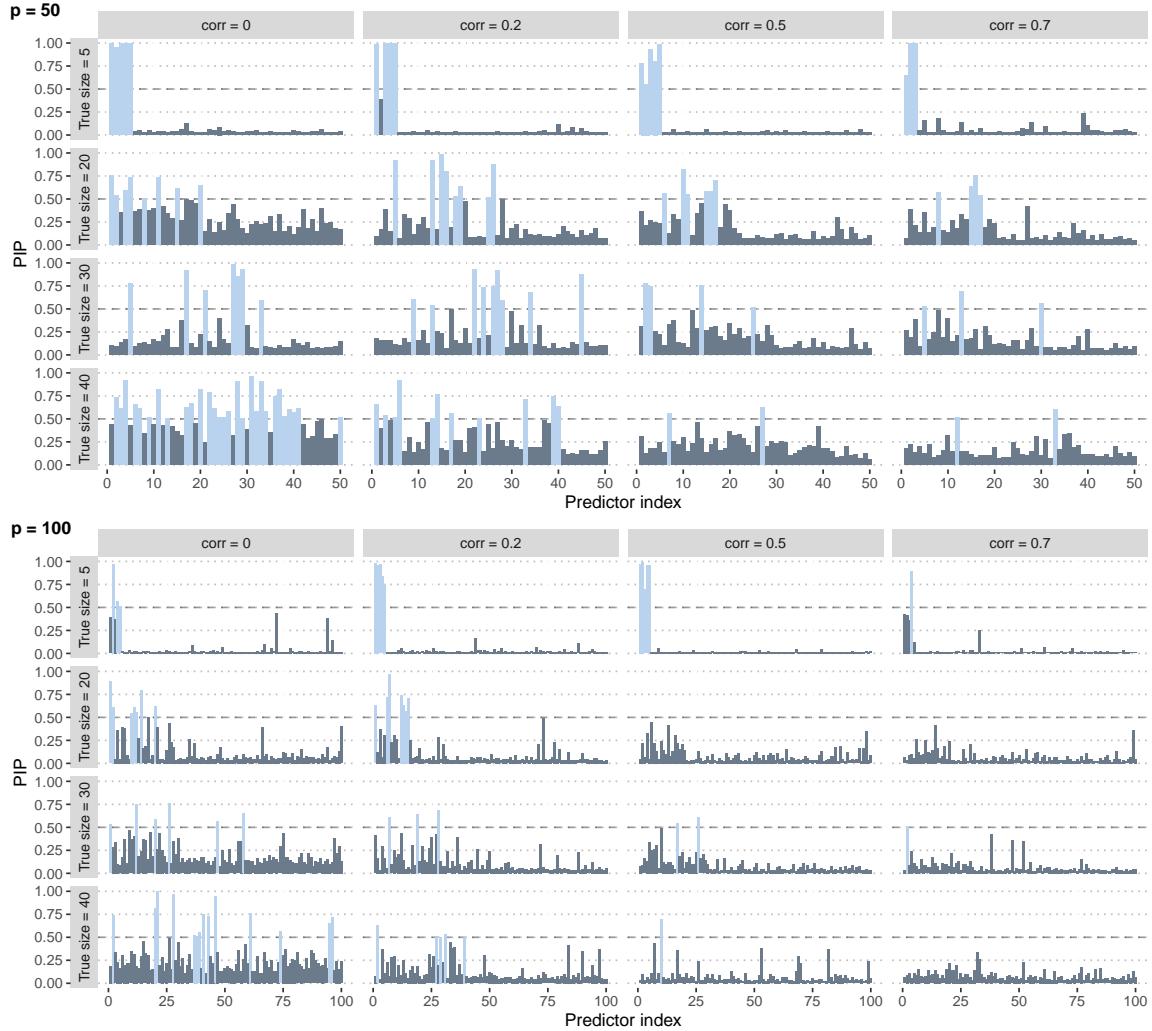
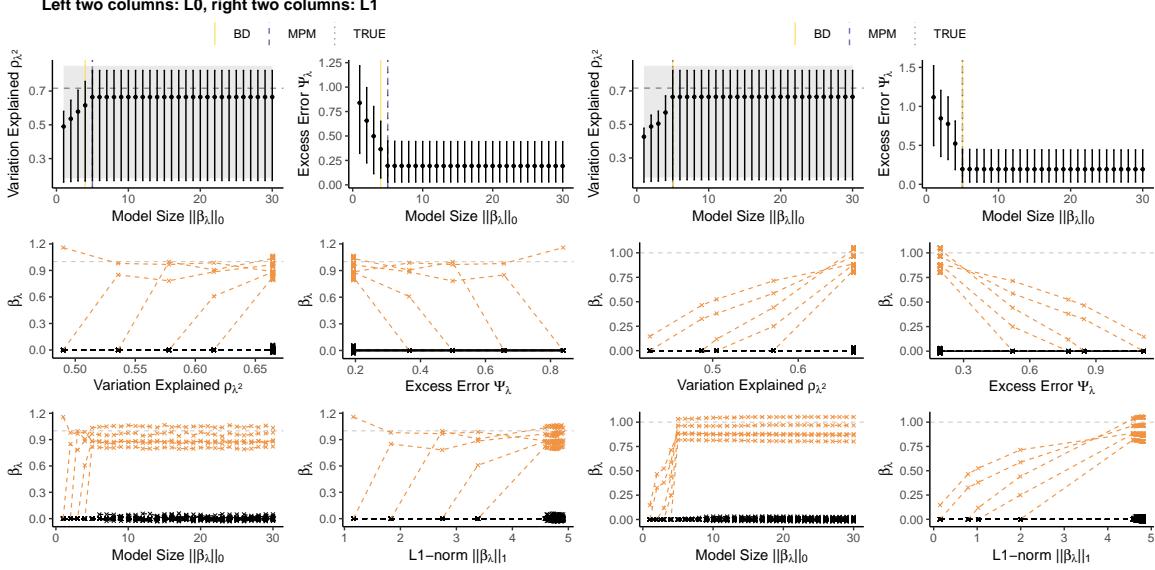


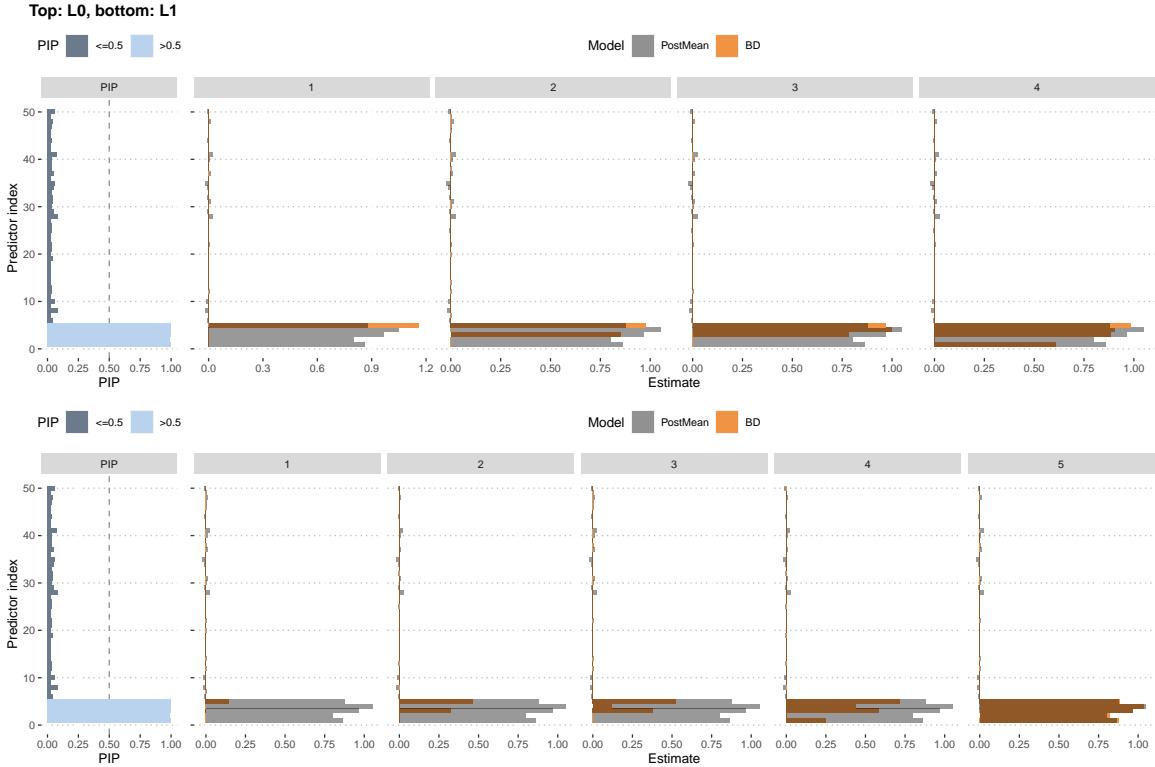
Figure 6: Posterior inclusion probabilities over all predictors on a single draw of the data, where  $n = 50$

## B.2 Bayesian decoupling selection summary plots in different simulations

### B.2.1 True model size = 5, correlation = 0



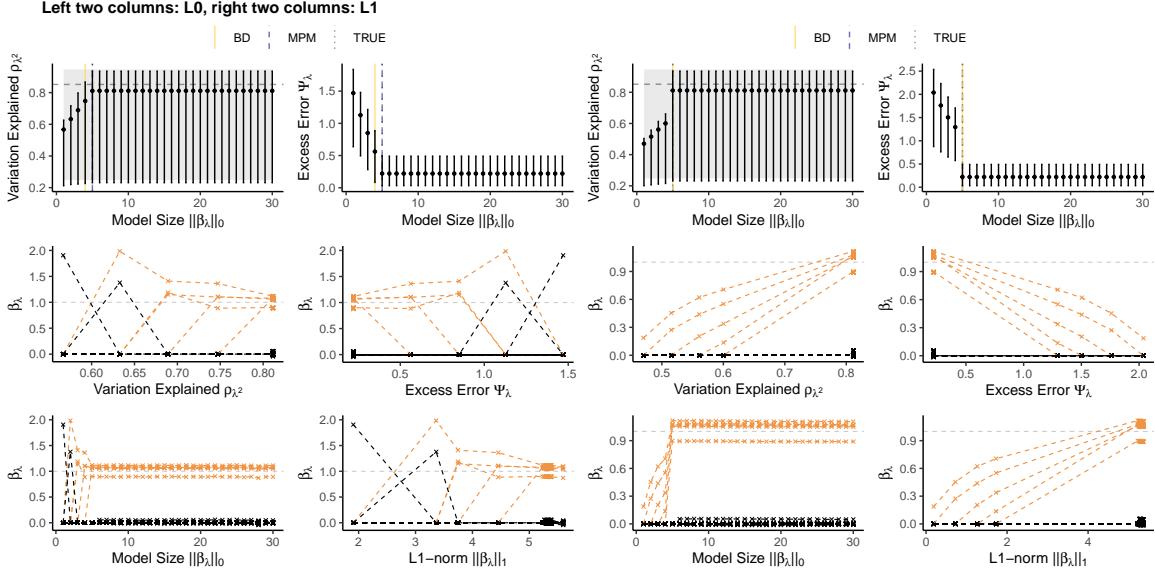
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



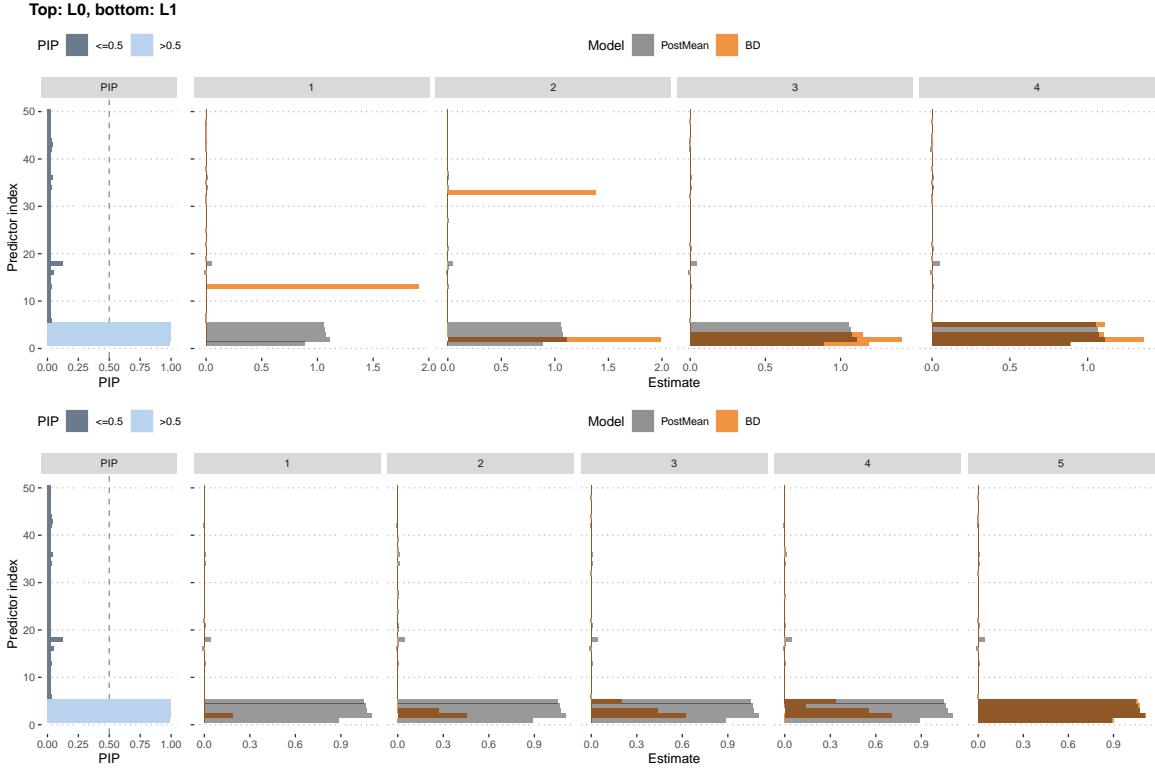
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 7: Bayesian decoupling solution paths when true model size = 5, correlation = 0, n = p = 50

### B.2.2 True model size = 5, correlation = 0.2



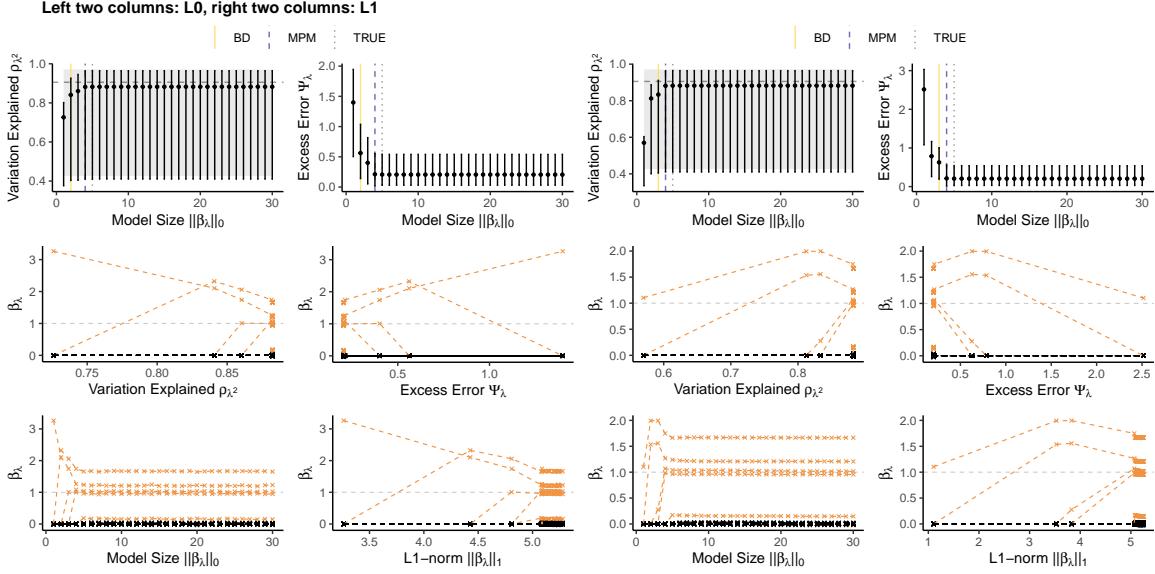
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



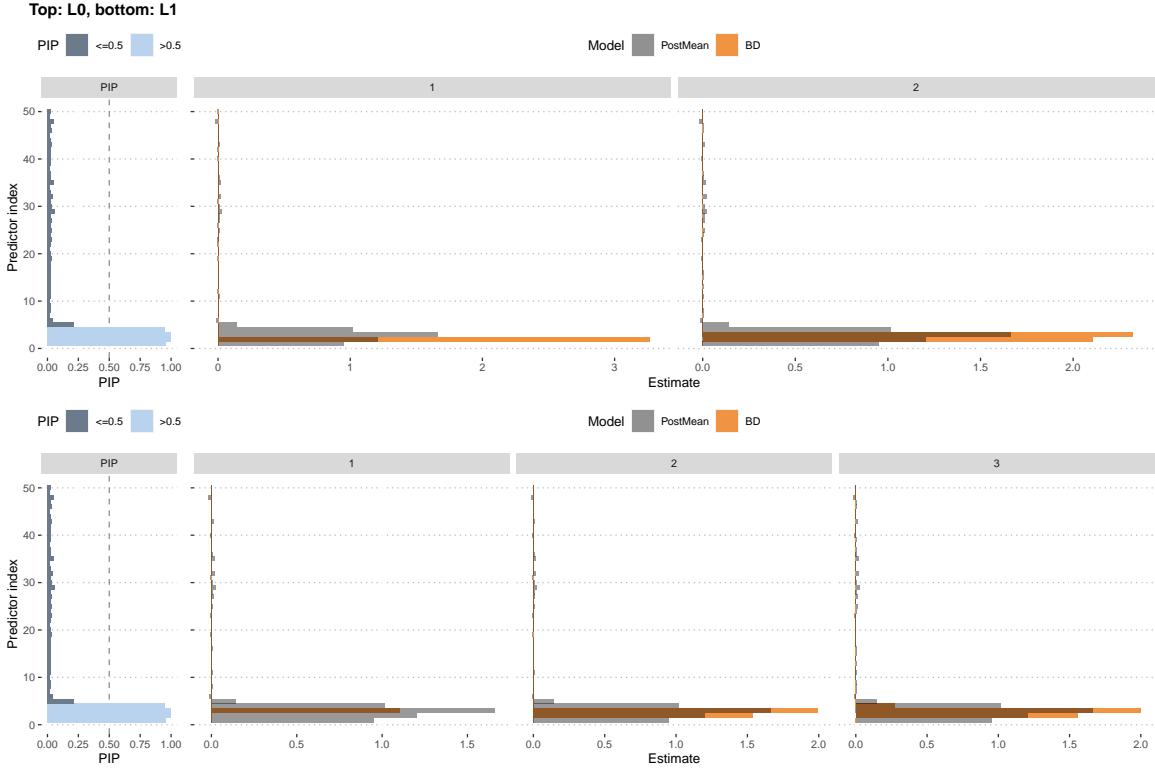
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 8: Bayesian decoupling solution paths when true model size = 5, correlation = 0.2, n = p = 50

### B.2.3 True model size = 5, correlation = 0.5



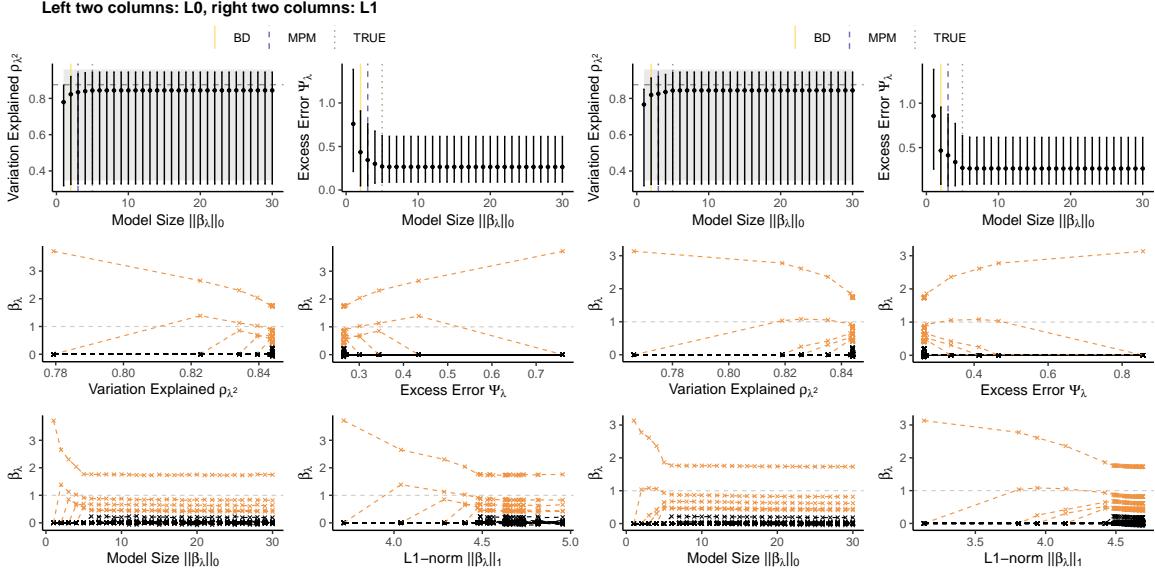
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\hat{\beta}_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



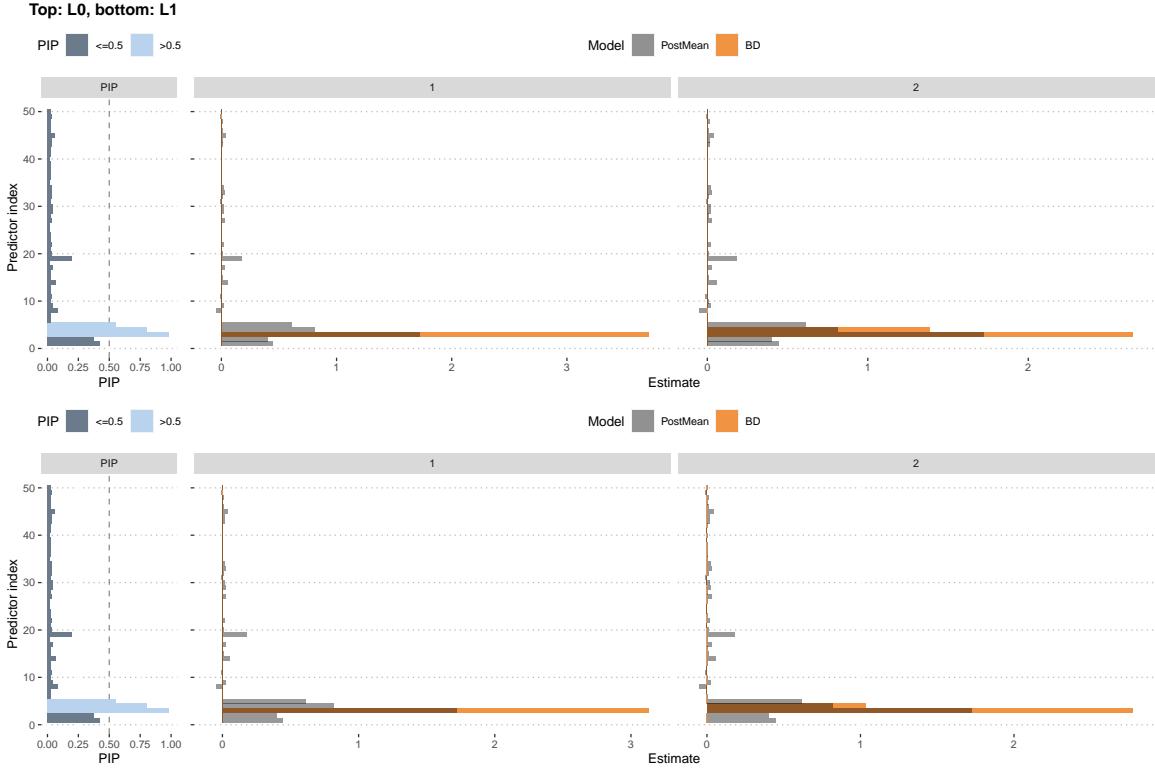
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 9: Bayesian decoupling solution paths when true model size = 5, correlation = 0.5, n = p = 50

### B.2.4 True model size = 5, correlation = 0.7



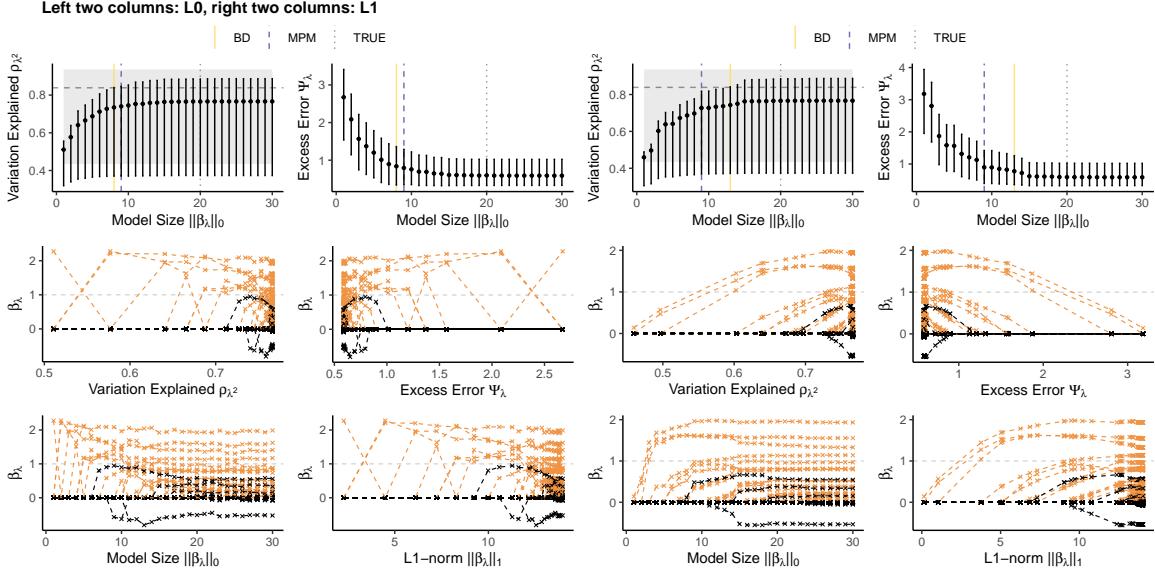
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



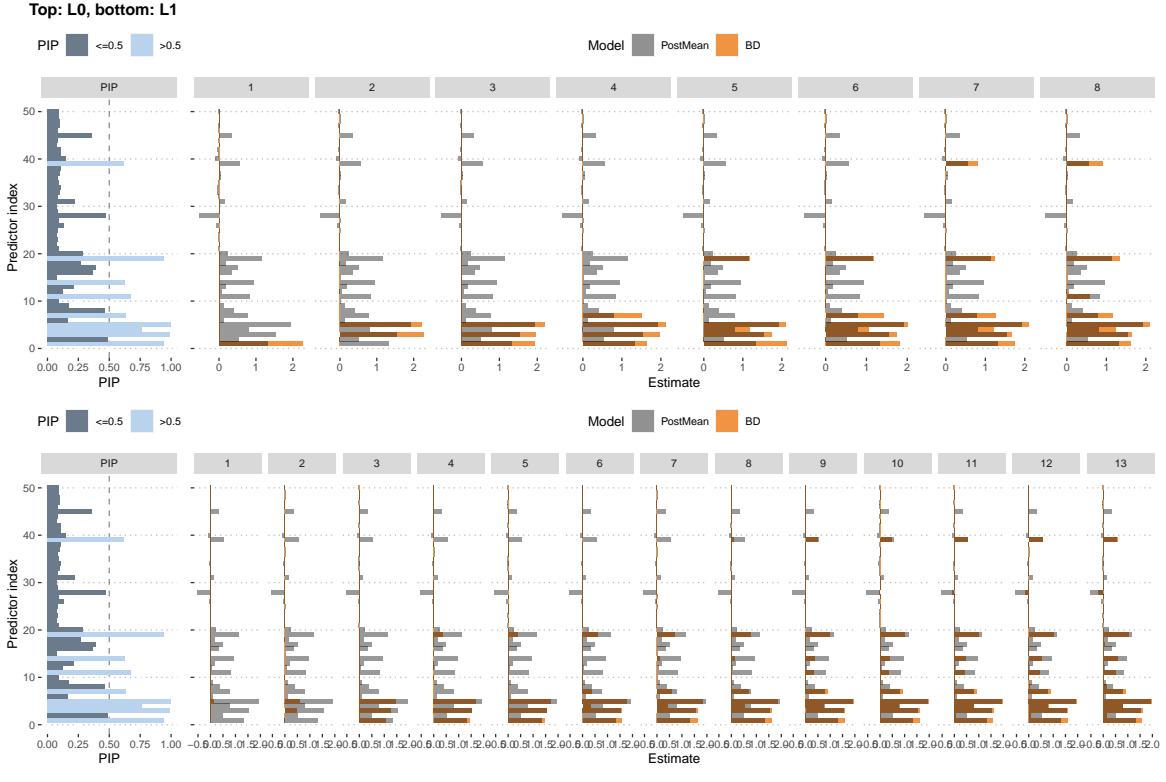
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 10: Bayesian decoupling solution paths when true model size = 5, correlation = 0.7, n = p = 50

### B.2.5 True model size = 20, correlation = 0



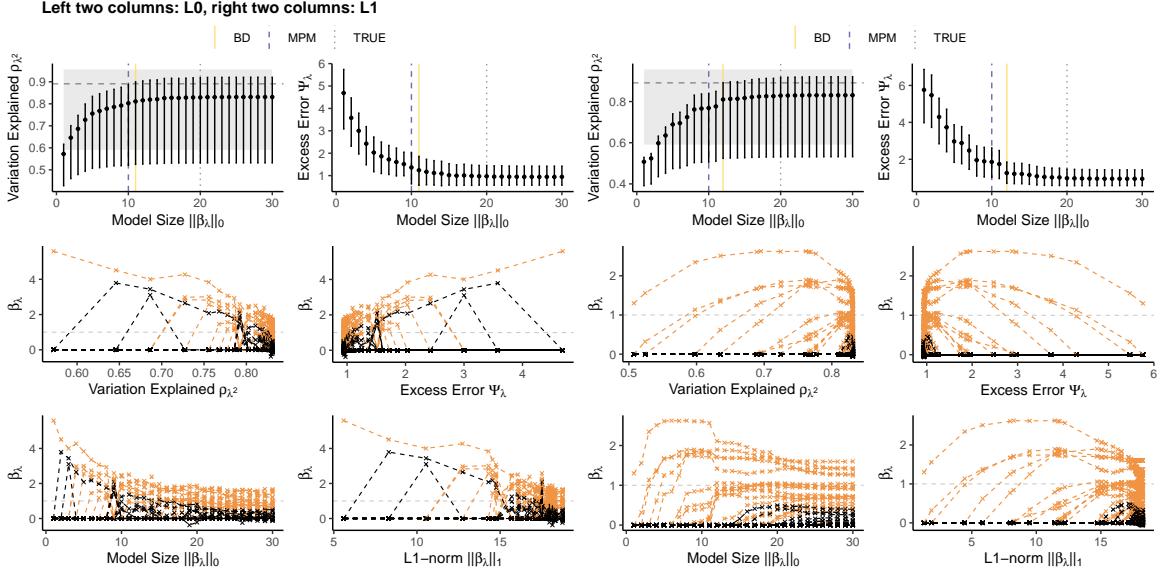
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



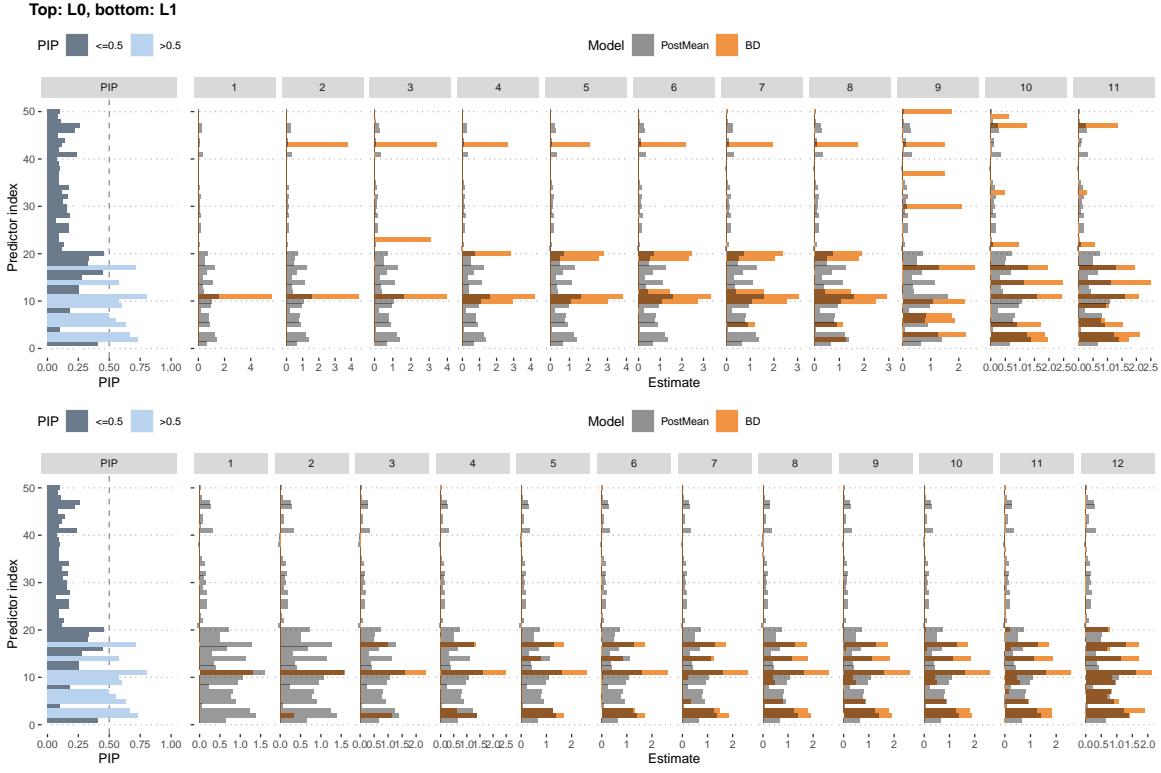
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 11: Bayesian decoupling solution paths when true model size = 20, correlation = 0,  $n = p = 50$

### B.2.6 True model size = 20, correlation = 0.2



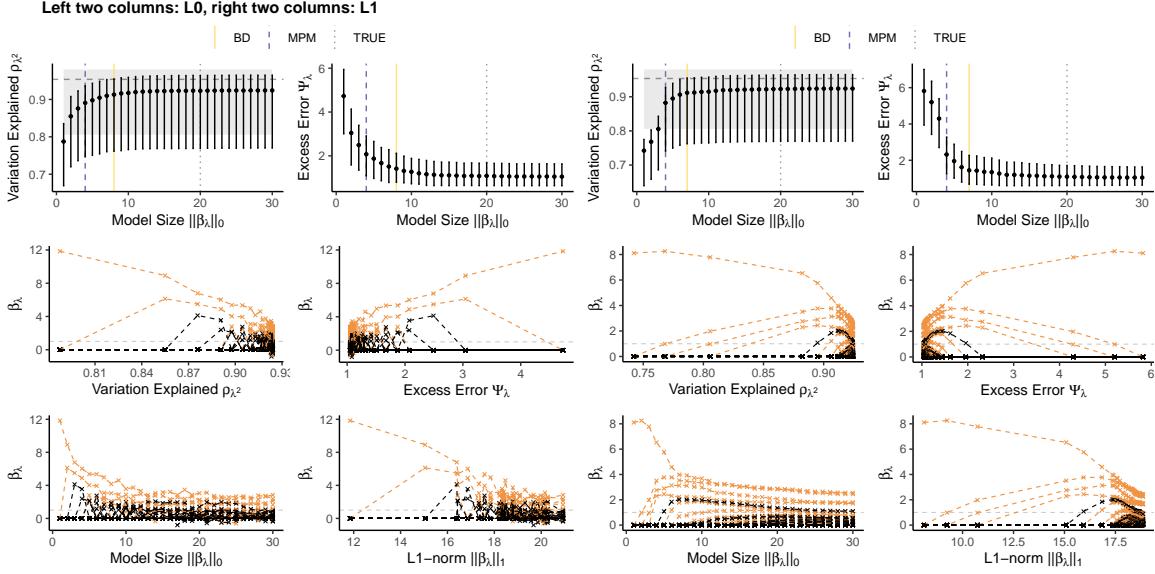
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



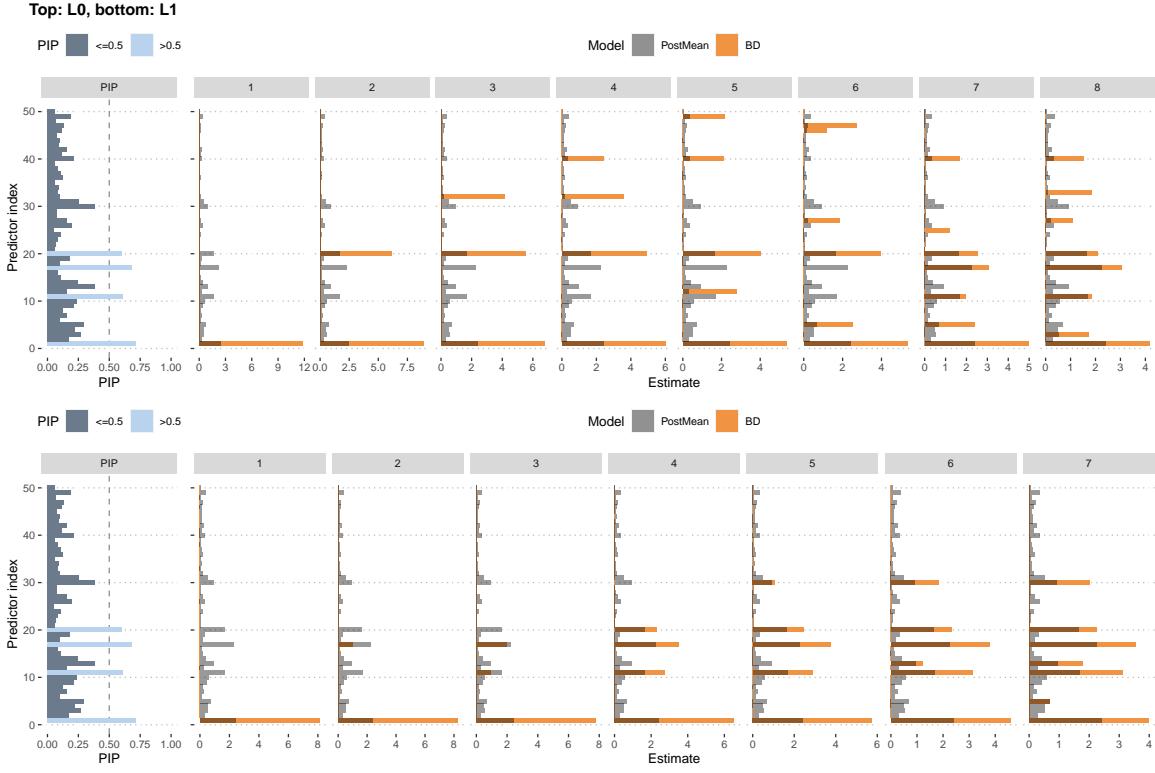
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 12: Bayesian decoupling solution paths when true model size = 20, correlation = 0.2, n = p = 50

### B.2.7 True model size = 20, correlation = 0.5



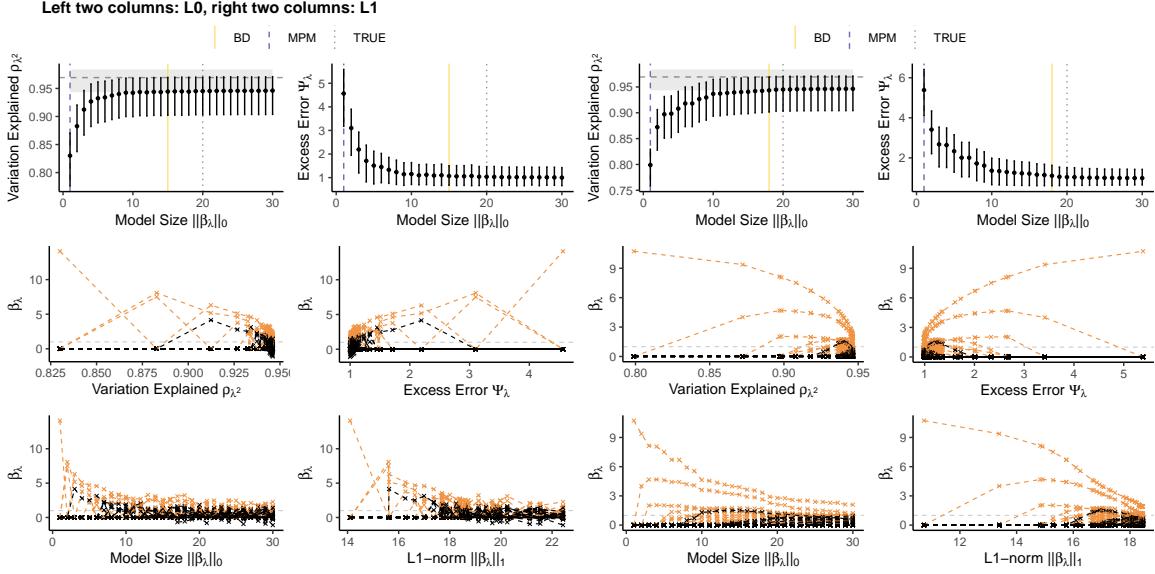
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



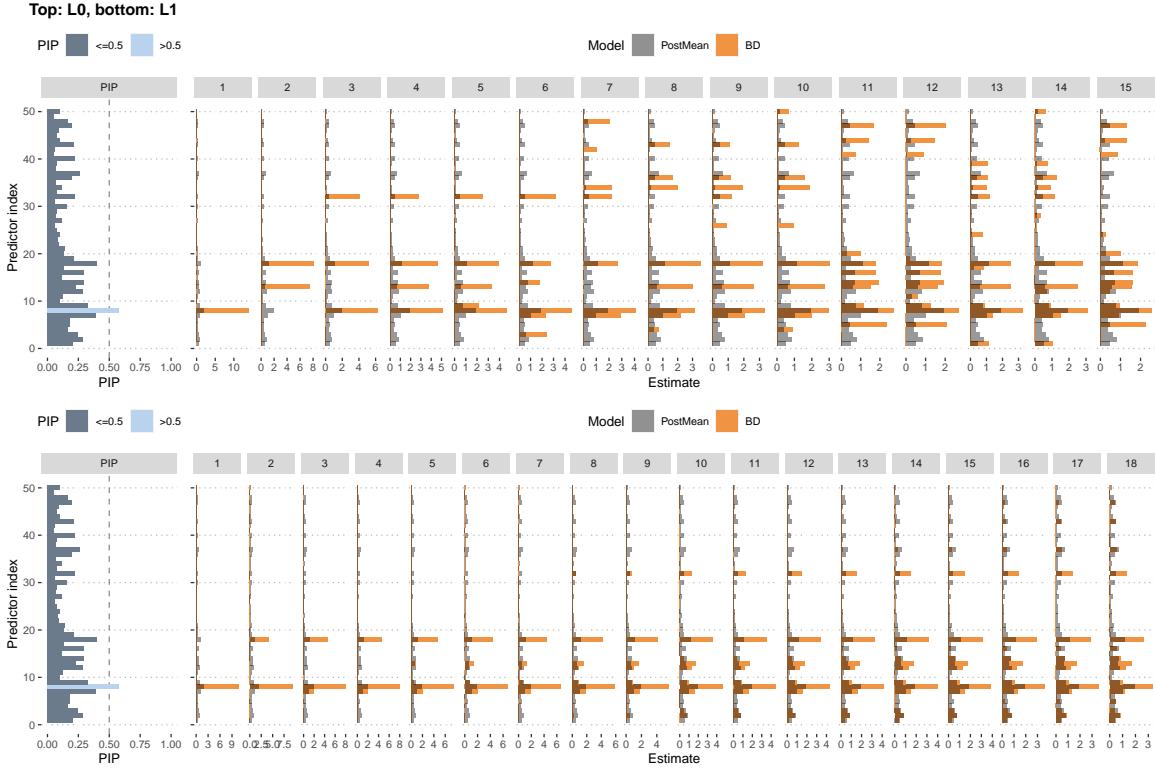
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 13: Bayesian decoupling solution paths when true model size = 20, correlation = 0.5, n = p = 50

### B.2.8 True model size = 20, correlation = 0.7



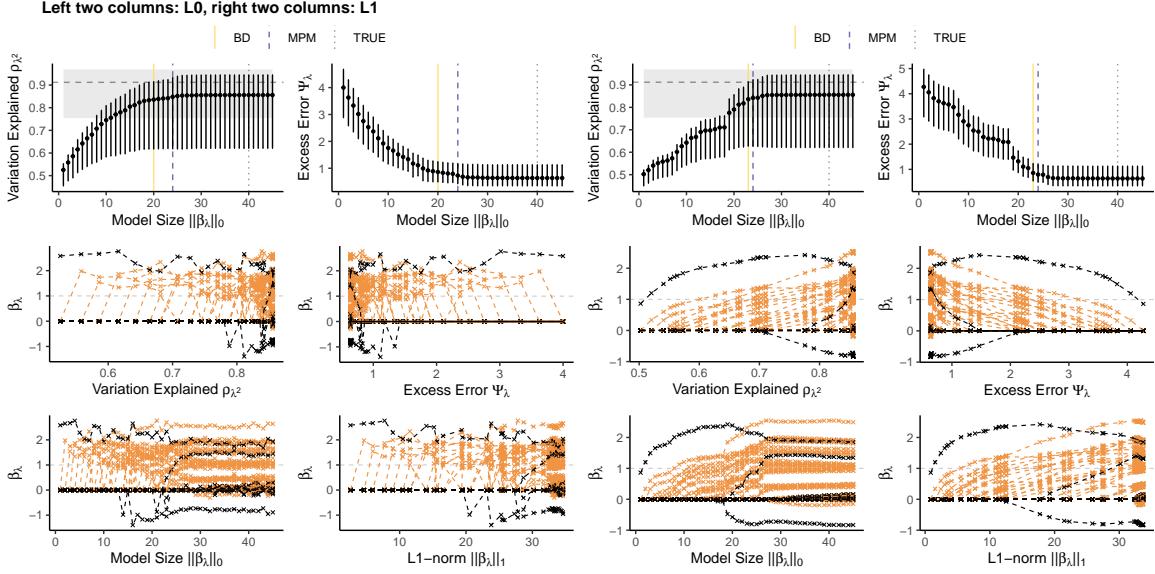
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



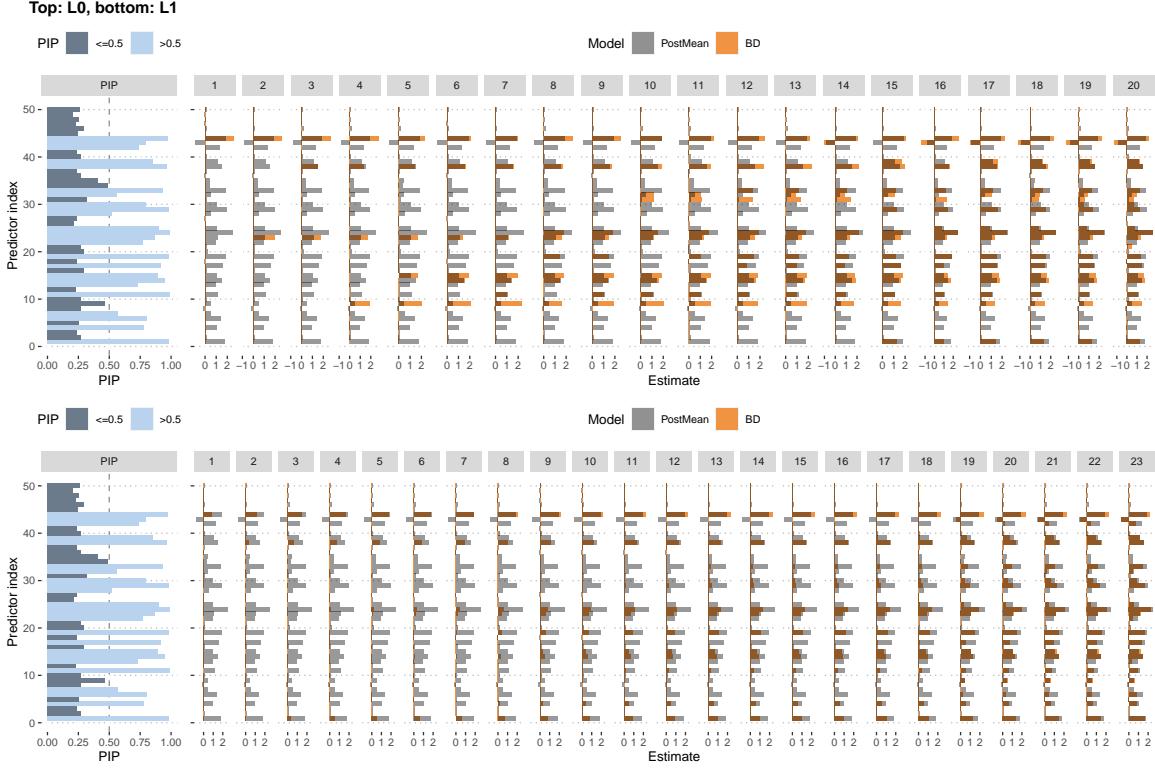
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 14: Bayesian decoupling solution paths when true model size = 20, correlation = 0.7, n = p = 50

### B.2.9 True model size = 40, correlation = 0



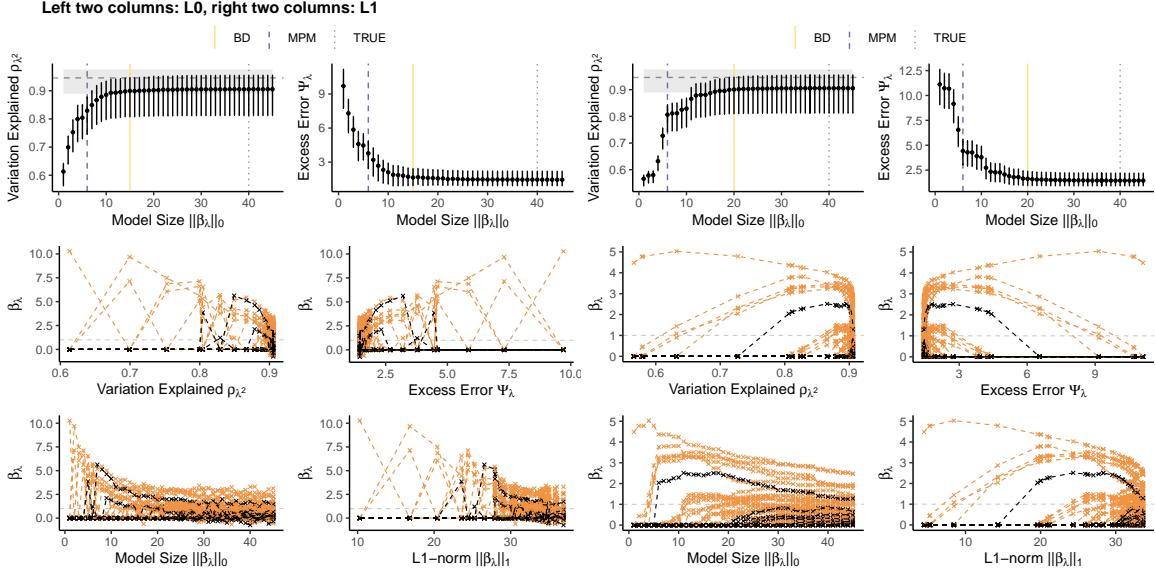
(a) Solution paths of all metrics. In the paths of  $p_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



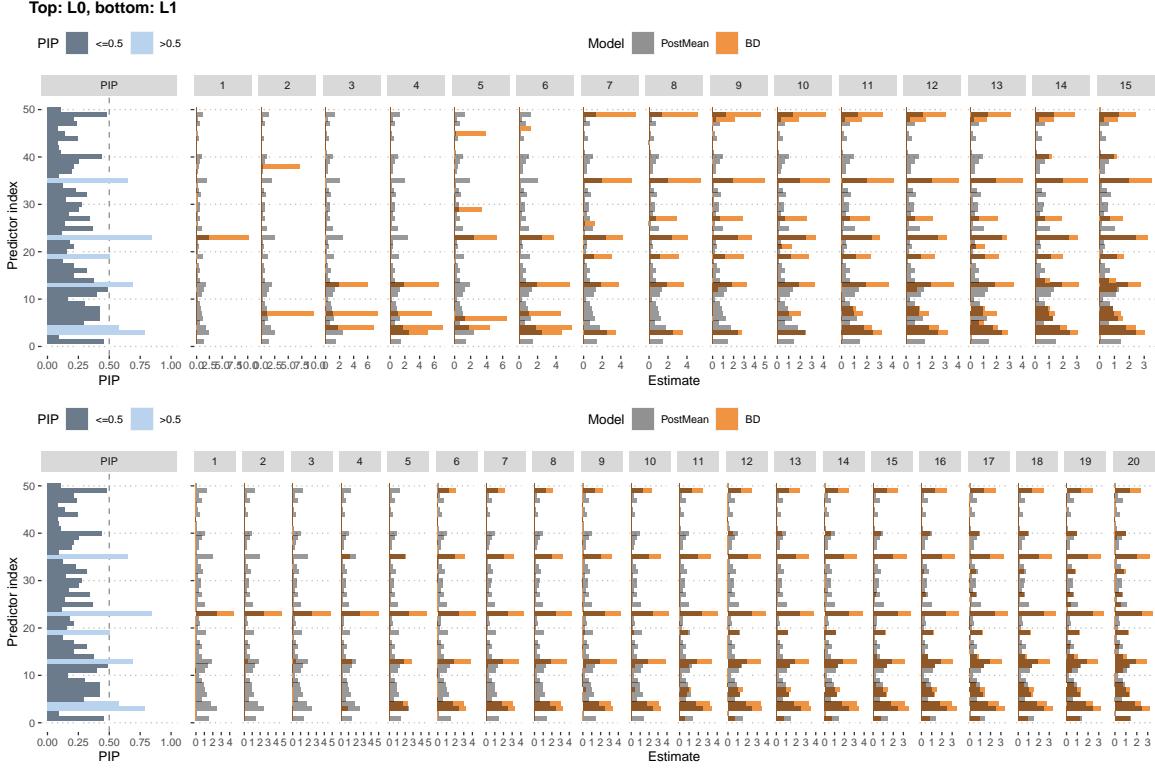
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 15: Bayesian decoupling solution paths when true model size = 40, correlation = 0,  $n = p = 50$

### B.2.10 True model size = 40, correlation = 0.2



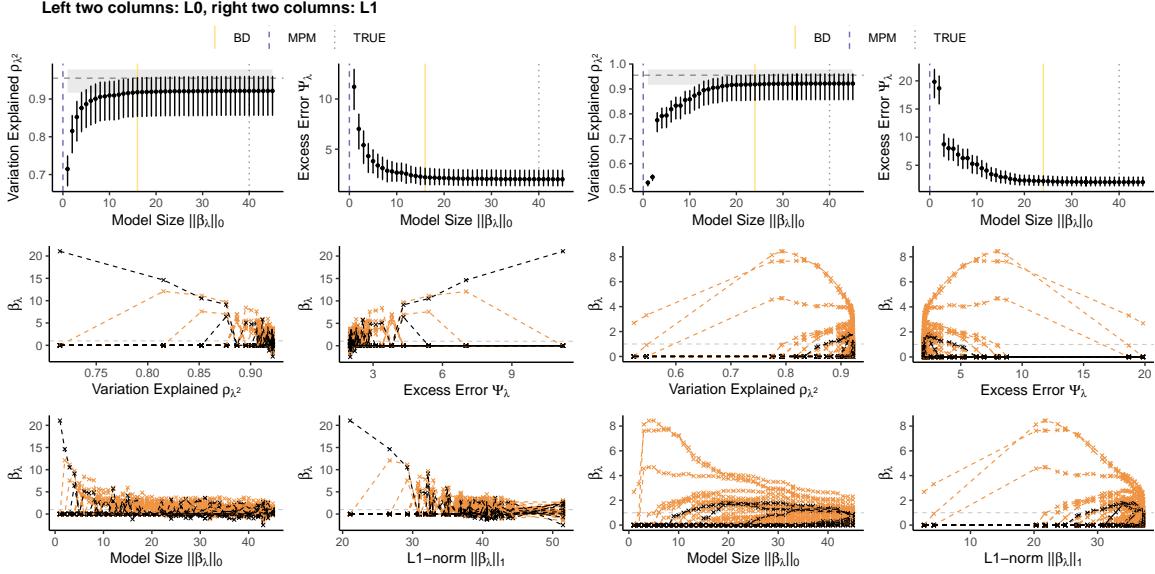
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



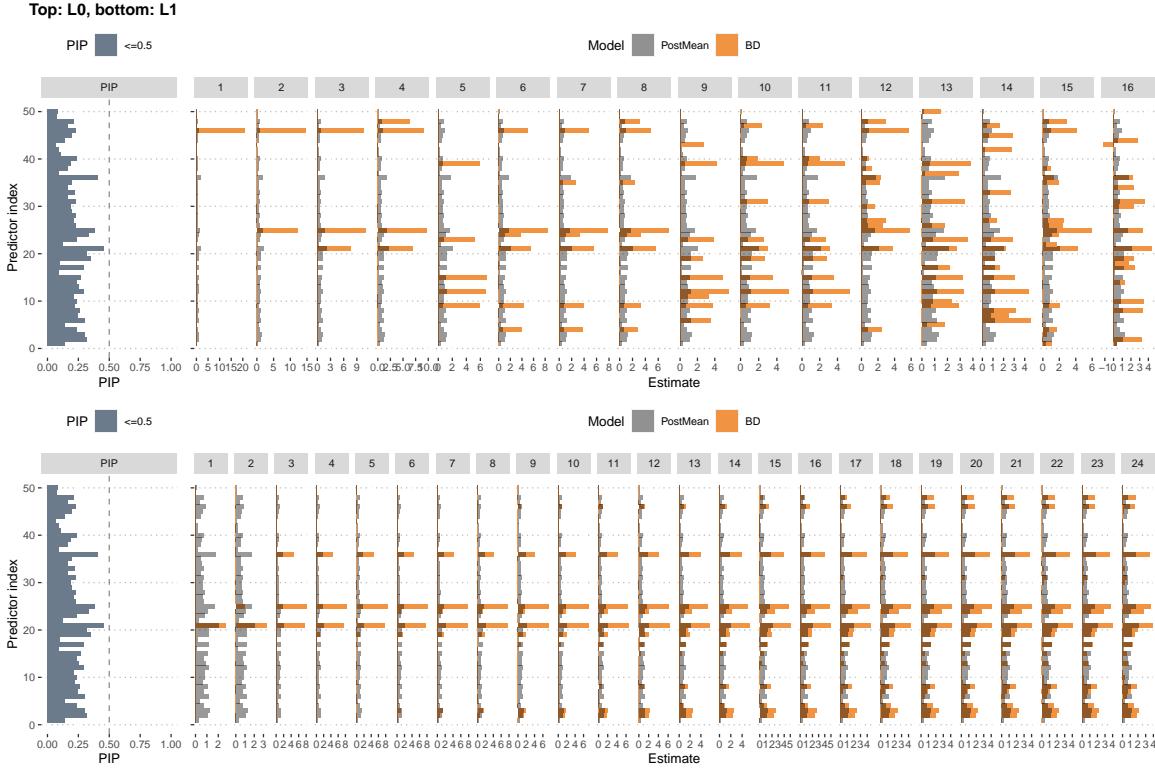
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 16: Bayesian decoupling solution paths when true model size = 40, correlation = 0.2, n = p = 50

### B.2.11 True model size = 40, correlation = 0.5



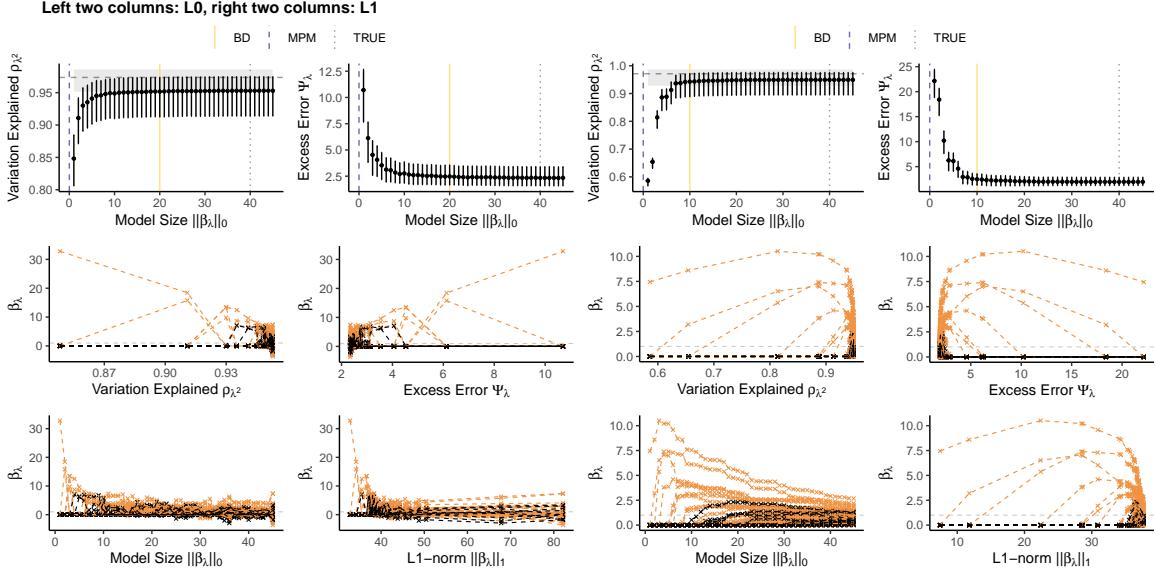
(a) Solution paths of all metrics. In the paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In the paths of  $\beta_\lambda$ , the yellow lines are non-zero signals, and the black lines are zero signals.



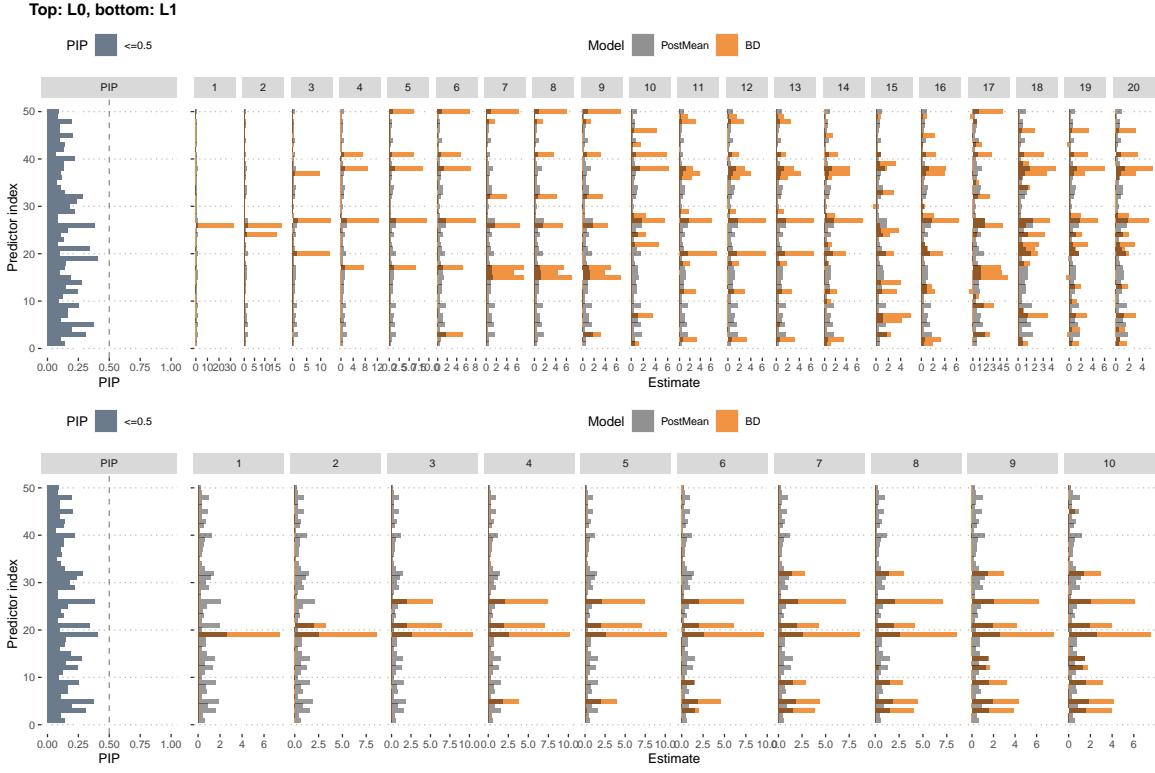
(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 17: Bayesian decoupling solution paths when true model size = 40, correlation = 0.5, n = p = 50

### B.2.12 True model size = 40, correlation = 0.7



(a) Solution paths of all metrics. In paths of  $\rho_\lambda^2$  and  $\psi_\lambda$ , the vertical lines represent the model sizes. In paths of  $\beta_\lambda$ , yellow lines are non-zero signals, black lines are zero signals.



(b) Decoupling solution versus posterior means and the posterior inclusion probabilities

Figure 18: Bayesian decoupling solution paths when true model size = 40, correlation = 0.7, n = p = 50