Statistics and Its Interface Volume 14 (2021) 211–227

Grouped variable selection with prior information via the prior group bridge method

Kai Li, Meng Mei, and Yuan Jiang[*],[†]

In a multiple regression with grouped predictors, it is usually desired to select important groups as well as to select important variables within a group simultaneously. To achieve this so-called "bi-level selection," group bridge has been developed as a combination of group-level bridge and variable-level lasso penalties. However, in many scientific areas, prior knowledge is available about the importance of certain groups of predictors, leading to the necessity of methodological development to incorporate such valuable information. For a prior-informative group, we propose a new penalty called "group ridge" as a combination of group-level ridge and variable-level lasso penalties, which always preserves this group while selects important variables in it. Then, we propose a composite group penalization named "prior group bridge" by applying group ridge and group bridge to prior-informative groups and groups with no prior information, respectively. We prove that prior group bridge achieves estimation and group selection consistencies given that the prior information is correct. In addition, we demonstrate the empirical advantage of prior group bridge over group bridge in terms of estimation, group and variable selection, and prediction through simulation studies. Finally, we apply prior group bridge to a genetic association study of bipolar disorder to illustrate its applicability and efficacy in real applications.

Keywords and phrases: Composite penalization, Group ridge, Selection consistency, Solution path.

1. INTRODUCTION

In many scientific problems, explanatory variables are naturally grouped. For example, in analysis of variance problems, a factor may have several levels and can be expressed through a group of dummy variables; in genetic association studies, single nucleotide polymorphisms (SNPs) are grouped in a gene and genes are grouped in a gene pathway. In these problems, it is often desired to select important individual variables and/or groups. Multiple statistical approaches have been proposed for this purpose, including group lasso [27, 15], group bridge [10], group SCAD, group MCP, cMCP [9], sparse-group lasso [20], the Composite Absolute Penalties (CAP) family [29], group exponential lasso [3], adaptive group lasso [24, 28, 25], doubly sparse penalty [13], group SLOPE [4], etc. We refer to Huang, Breheny and Ma [9] for a review of grouped variable selection methods.

The above-mentioned methods can be categorized into two types: those that only carry out group-level selection and those that carry out bi-level selection, i.e., selection at both group and individual levels. Group-level selection methods only select important groups; within a group, the coefficients of individual variables are forced to be either all zero or all nonzero. By contrast, bi-level selection methods can select important groups as well as useful variables within those groups. Among the above-mentioned methods, group lasso, group MCP, group SCAD, adaptive group lasso, and group SLOPE are all group-level selection methods, while group bridge, cMCP, sparse-group lasso, group exponential lasso, and doubly sparse penalty are bi-level selection methods. The CAP family in Zhao et al. [29] includes group lasso and group bridge as its special cases and is more general than the other methods.

In practice, prior knowledge is sometimes available about the importance of a certain group of predictors but not individual ones. For example, school district as a factor is well known to be associated with house price. However, one might still be interested in telling whether the house prices in two particular school districts have a significant difference or not. In this case, the prior knowledge is about the importance of the school district as a whole factor but not necessarily all its levels (compared to the baseline). Another example arises in genetic association studies. When an association study is conducted to identify genetic signals at the SNP level, a risk gene may have been discovered by prior studies. In this case, the prior knowledge is about the importance of a gene, i.e., a group of SNPs, instead of all the SNPs in this gene. In both examples, the prior knowledge can be represented as a prior-informative group of predictors, such as the factor of all school districts or the SNPs in the risk gene, for the new studies.

Another reason for the necessity of incorporating the prior knowledge is the difficulty detecting signals from the massive data collected by current research. For example, in a typical genetic association study, hundreds of thousands of SNPs are genotyped, and yet their individual effect sizes are usually very small. It is very unlikely to detect all important genetic factors with a desired statistical significance in

[*] Corresponding author.

[†] ORCID: 0000-0001-6409-9159.

a single study; therefore, top findings from various studies do not usually show obvious overlaps [14]. This is referred to as the notorious "lack of reproducibility" problem. Therefore, it is imperative for us to borrow information from other studies, such as the known risk SNPs or genes, in order to improve reproducibility in genetic findings. Jiang, He and Zhang [11] proposed a method called "prior lasso" to incorporate prior-informative variables into a variable selection problem. However, to the best of our knowledge, no method is available yet to incorporate prior-informative groups into a (grouped) variable selection problem. This motivates us to develop such a method.

For a prior-informative group, one expects it to be also important in a new study. However, we would also like to select important variables instead of including all variables in it. In the house price example, it is totally possible that the house prices in two particular school districts do not differ significantly; in the genetic association study, it is usually the case that only a small number of SNPs in a risk gene are associated with the disease. In other words, not only do we want to preserve a prior-informative group, but we would also like to select individual variables within the group. It is noteworthy that none of the above-mentioned grouped variable selection methods can achieve this goal. Applying either group-level or bi-level selection methods, one may still lose the whole prior-informative group because none of these methods guarantee to retain a group in the model.

To fulfill our purpose, we propose a new penalty named "group ridge" which can preserve a group while performing variable selection within it. We show that at least one of the coefficients of the variables within a group is nonzero under this new penalty. In addition, we apply group ridge and group bridge to prior-informative groups and other groups without any prior information, respectively, leading to a novel composite group penalization named "prior group bridge." Compared to other grouped variable selection methods, prior group bridge treats prior-informative groups differently from the other groups. If correct prior information is incorporated, prior group bridge takes the advantage and outperforms the other methods.

The rest of the paper is organized as follows. In Section 2, we will introduce group ridge and investigate its theoretical properties with comparison to other penalization methods. In Section 3, we will introduce prior group bridge as a composite group penalization combining group ridge and group bridge. In addition, we will establish the asymptotic theory and develop an efficient algorithm for prior group bridge. In Section 4, we will show the empirical advantage of prior group bridge over group bridge through simulation studies. Prior group bridge works better than group bridge especially for those groups with weak signals, which are common in many real problems. In Section 5, we apply prior group bridge to a real genetic association study of bipolar disorder. Compared to group bridge, it leads to findings that are more consistent with the knowledge from previous studies thus improves reproducibility. Section 6 concludes this paper with some discussion. Some details for the theoretical proof in Sections 2 and 3 are given in the Appendix.

2. GROUP RIDGE

2.1 Group lasso and group bridge

Suppose we have a set of independent and identically distributed observations $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^\top$ is a d-dimensional vector of covariates and Y_i is the observed response given the corresponding $\mathbf{X}_i$. Let $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$ be the design matrix, $\mathbf{x}_k = (X_{1k}, \ldots, X_{nk})^\top$ be its kth column with $k = 1, \ldots, d$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ be the response vector. Assume that the conditional distribution of Y_i given $\mathbf{X}_i$ belongs to the canonical exponential family with the following density function:

$$f(Y_i|\theta_i) \propto \exp[Y_i\theta_i - b(\theta_i)], \tag{1}$$

where $\theta_i = \beta_0 + \mathbf{X}_i^\top\boldsymbol{\beta}$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top$ and $b(\cdot)$ is the canonical link function assumed to be twice continuously differentiable with a positive second-order derivative, $b''(\cdot) > 0$. The loss function of a generalized linear model is simply the negative log-likelihood function:

$$\begin{aligned} L(\beta_0, \boldsymbol{\beta}) &= -\sum_{i=1}^{n}[Y_i\theta_i - b(\theta_i)] \\ &= -\sum_{i=1}^{n}[Y_i(\beta_0 + \mathbf{X}_i^\top\boldsymbol{\beta}) - b(\beta_0 + \mathbf{X}_i^\top\boldsymbol{\beta})]. \end{aligned}$$

We impose a group structure on the covariates. Let $A_1, \ldots, A_J$ be a partition of the set $\{1, \ldots, d\}$ representing groups of the covariate vector and denote the regression coefficients in the group A_j by $\boldsymbol{\beta}_{A_j} = (\beta_k : k \in A_j)^\top$. In addition, for any m-dimensional vector $\mathbf{a} = (a_1, \ldots, a_m)^\top$, denote its L_1 norm by $\|\mathbf{a}\|_1 = |a_1| + \cdots + |a_m|$ and its L_2 norm by $\|\mathbf{a}\|_2 = (a_1^2 + \cdots + a_m^2)^{1/2}$.

Yuan and Lin [27] proposed group lasso for grouped variable selection, leading to the group lasso estimator defined as the minimizer of

$$\ell(\beta_0, \boldsymbol{\beta}) = L(\beta_0, \boldsymbol{\beta}) + \lambda\sum_{j=1}^{J} c_j\|\boldsymbol{\beta}_{A_j}\|_2, \tag{2}$$

where $\lambda > 0$ is a tuning parameter and c_j's are constants to adjust the different dimensions of $\boldsymbol{\beta}_{A_j}$. A simple choice is $c_j \propto |A_j|^{1/2}$, where $|A_j|$ is the cardinality of A_j.

A unique characteristic of group lasso is that it selects important variables at the group level, but not at the individual variable level [27]. To achieve both group and individual variable selections (the so-called bi-level selection), Huang et al. [10] proposed the group bridge estimator that minimizes the following objective function:

$$(3) \qquad \ell(\beta_0, \boldsymbol{\beta}) = L(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^{\gamma}, \ \gamma \in (0,1),$$

where $\lambda > 0$ and c_j's have a similar role to those in (2). Huang et al. [10] chose $c_j \propto |A_j|^{1-\gamma}$. The group bridge method can be used for variable selection at the group and individual variable levels simultaneously.

2.2 Group ridge

In our applications, prior knowledge is available about the importance of certain groups. For such a prior-informative group, we would like to preserve the group in the model while to select important variables within this group. Neither group lasso nor group bridge can achieve this goal as both penalties might exclude the whole group from the model.

For our purpose, we propose a new penalty in conjunction with the same loss function in (2) and (3) as follows:

$$(4) \qquad \ell(\beta_0, \boldsymbol{\beta}) = L(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^{2},$$

where $\lambda > 0$ and c_j's have a similar role to those in (2) and (3). Inspired by the name of group bridge, we call the penalty $\|\boldsymbol{\beta}_{A_j}\|_1^2$ the group ridge penalty. Group ridge can be regarded as a combination of group-level ridge and variable-level lasso penalties.

To illustrate the statistical property as well as to derive the computational algorithm for group ridge, let us first present the following Karush-Kuhn-Tucker (KKT) conditions for minimizing (4).

Theorem 1. *A set of necessary and sufficient conditions for an estimator $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ to be a global solution of (4) is that, for all $j = 1, \ldots, J$,*

$$(5) \qquad \sum_{i=1}^{n} [Y_i - b'(\hat{\beta}_0 + \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}})] = 0,$$

$$(6) \qquad \left| \sum_{i=1}^{n} [Y_i - b'(\hat{\beta}_0 + \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}})] X_{ik} \right| \le 2\lambda c_j \|\hat{\boldsymbol{\beta}}_{A_j}\|_1, \quad \textit{when } \hat{\beta}_k = 0, \ k \in A_j,$$

$$(7) \qquad \sum_{i=1}^{n} [Y_i - b'(\hat{\beta}_0 + \mathbf{X}_i^{\top} \hat{\boldsymbol{\beta}})] X_{ik} = 2\lambda c_j \|\hat{\boldsymbol{\beta}}_{A_j}\|_1 \operatorname{sign}(\hat{\beta}_k), \quad \textit{when } \hat{\beta}_k \neq 0, \ k \in A_j.$$

The KKT conditions for a lasso solution [23] are very similar to those for group ridge in (5)–(7), except that they do not have the term $\|\hat{\boldsymbol{\beta}}_{A_j}\|_1$ on the right-hand sides of (6) and (7). This small but key distinction provides us with some intuitions about how group ridge differs from lasso as follows.

For lasso, if (6) is satisfied without $\|\hat{\boldsymbol{\beta}}_{A_j}\|_1$ on the right-hand side for all $k \in A_j$, then the estimators of the whole group $\hat{\boldsymbol{\beta}}_{A_j}$ will be shrunk to zero. However, for group ridge, it is always true that $\hat{\boldsymbol{\beta}}_{A_j} \neq \mathbf{0}$ for all $j = 1, \ldots, J$. Otherwise, $\|\hat{\boldsymbol{\beta}}_{A_j}\|_1$ becomes 0 and consequently the right-hand side of (6) becomes zero. This will result in an unpenalized estimator for group A_j as the first-order derivatives [left-hand side of (6)] are all zeros, which contradicts with the fact that $\hat{\boldsymbol{\beta}}_{A_j} = \mathbf{0}$.

In addition to the above intuitive arguments, a more rigorous proof will be provided in Section 2.3 to show that group ridge will keep at least one nonzero coefficient in any group. It is a unique characteristic not possessed by lasso, group lasso, or group bridge. This property meets our need to preserve prior-informative groups from existing knowledge while to select important variables within these groups.

2.3 Solution path

For linear regression, the development of the lars algorithm [7] leads to the geometric interpretation of the lasso estimator, and consequently, a clear illustration of how the lasso estimator evolves when the tuning parameter changes. We notice that group ridge is simply the square of the lasso penalty if there is only one group. Therefore, we will develop a similar solution path for group ridge on one group and compare it with lars to better characterize the geometric properties of the group ridge estimator.

In the one-group setting, let's assume the generalized linear model in (1) with $\theta_i = \beta_0 + \mathbf{X}_i^{\top} \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\top}$ is in a single group. In this case, the group ridge estimator $[\hat{\boldsymbol{\beta}}_0(\lambda), \hat{\boldsymbol{\beta}}(\lambda)]$ is defined to be the minimizer of

$$(8) \qquad \ell(\beta_0, \boldsymbol{\beta}) = L(\beta_0, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1^2.$$

Before presenting the group ridge solution path, we point out an interesting observation that links the group ridge solution to the lasso solution. Define the corresponding lasso solution, $[\hat{\alpha}_0(\mu), \hat{\boldsymbol{\alpha}}(\mu)]$, to be the minimizer of

$$(9) \qquad \ell(\alpha_0, \boldsymbol{\alpha}) = L(\alpha_0, \boldsymbol{\alpha}) + \mu \|\boldsymbol{\alpha}\|_1.$$

Simply speaking, there is a one-to-one correspondence between the two solutions. We summarize this result in the following theorem.

Theorem 2. *Let the group ridge solution be $[\hat{\boldsymbol{\beta}}_0(\lambda), \hat{\boldsymbol{\beta}}(\lambda)]$ for (8). Define $\mu = 2\lambda \|\hat{\boldsymbol{\beta}}(\lambda)\|_1$. Then, the lasso solution for (9), $[\hat{\alpha}_0(\mu), \hat{\boldsymbol{\alpha}}(\mu)]$, is equal to $[\hat{\boldsymbol{\beta}}_0(\lambda), \hat{\boldsymbol{\beta}}(\lambda)]$.*

Theorem 2 can be easily proved by comparing the KKT conditions for both solutions. It illustrates a one-to-one correspondence between the group ridge and lasso solutions. However, this result is less helpful in finding one solution from the other due to the lack of direct relationship between μ and λ. For example, suppose we know the lasso solution path, i.e., the lasso solutions $[\hat{\alpha}_0(\mu), \hat{\boldsymbol{\alpha}}(\mu)]$ for all $\mu > 0$. Then, for a given $\lambda > 0$, can we easily find $[\hat{\boldsymbol{\beta}}_0(\lambda), \hat{\boldsymbol{\beta}}(\lambda)]$ using Theorem 2? Unfortunately, it is not straightforward because there is no easy way to determine the corresponding μ for this particular λ through Theorem 2.

Therefore, we will develop the solution path for group ridge as follows, which also provides a deeper geometric insight than Theorem 2. For simplicity, we assume the "one at a time" assumption as in Efron et al. [7]. Let $\mathcal{A}(\lambda)$ denote the active set of the estimator $\hat{\boldsymbol{\beta}}(\lambda)$, i.e., $\mathcal{A}(\lambda) = \{k \in \{1, \ldots, d\} : \hat{\beta}_k(\lambda) \neq 0\}$. We start with a very large λ at the beginning of the path. It is well known that the lasso estimator $\hat{\boldsymbol{\alpha}}(\mu)$ (except the intercept) will be shrunk to zero when μ is large enough. However, our result suggests that group ridge has a nonzero starting point, different from lasso.

Theorem 3. *When λ is large enough, $\mathcal{A}(\lambda)$ of the group ridge estimator $\hat{\boldsymbol{\beta}}(\lambda)$ always includes one index k^*.*

Obviously, no matter how large λ is, group ridge will keep at least one nonzero coefficient. It is why group ridge can preserve prior-informative groups compared to other penalties. In addition, it is easy to find k^* under specific models. For example, consider linear regression in which $b(\theta) = \theta^2/2$ in (1). Further, suppose that both $\mathbf{Y}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are centralized so that the intercept β_0 is not included in (8). In this case, k^* satisfies that $|\mathbf{x}_{k^*}^\top \mathbf{Y}| = \max_{1 \le k \le d} |\mathbf{x}_k^\top \mathbf{Y}|$. Meanwhile, $\hat{\beta}_{k^*}(\lambda) = \mathbf{x}_{k^*}^\top \mathbf{Y} / (\mathbf{x}_{k^*}^\top \mathbf{x}_{k^*} + 2\lambda)$ when λ is large enough. In other words, the predictor with the largest absolute covariance with $\mathbf{Y}$ will always have a nonzero coefficient even when $\lambda \to \infty$.

As the tuning parameter λ decreases from ∞ to 0, the indices $k \in \{1, \ldots, d\}$ will either be added to or removed from $\mathcal{A}(\lambda)$, and when $\lambda = 0$, $\mathcal{A}(0)$ will match the active set of the unpenalized estimator $\hat{\boldsymbol{\beta}}(0)$. Thus, the solution path is built in consecutive steps, each step adding or deleting a covariate under the "one at a time" assumption [7]. Within a given step, the active set $\mathcal{A}(\lambda)$ stays unchanged. Therefore, to figure out the whole path, we need to clarify how $\hat{\boldsymbol{\beta}}(\lambda)$ evolves within a step as well as how $\mathcal{A}(\lambda)$ is updated between two consecutive steps. The next result shows explicitly how $\hat{\boldsymbol{\beta}}(\lambda)$ evolves within a step.

Theorem 4. *Within a step starting at λ_0, $\mathcal{A}(\lambda)$ stays the same as $\mathcal{A}(\lambda_0) = \mathcal{A}$, and the group ridge estimator $[\hat{\boldsymbol{\beta}}_0(\lambda), \hat{\boldsymbol{\beta}}(\lambda)]$ satisfies that*

$$\begin{aligned}(10)\quad & \mathbb{X}_\mathcal{A}^\top \{b'[\hat{\beta}_0(\lambda) + \mathbb{X}_\mathcal{A}^\top \hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda)] - b'[\hat{\beta}_0(\lambda_0) + \mathbb{X}_\mathcal{A}^\top \hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_0)]\} \\ & = -2\mathbf{S}_\mathcal{A}\mathbf{S}_\mathcal{A}^\top [\lambda \hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda) - \lambda_0 \hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_0)],\end{aligned}$$

where $\mathbb{X}_\mathcal{A} = (\ldots, \mathbf{x}_k, \ldots)_{k \in \mathcal{A}}$ and $\mathbf{S}_\mathcal{A} = (\ldots, S_k, \ldots)_{k \in \mathcal{A}}^\top$ with $S_k = \operatorname{sign}[\hat{\beta}_k(\lambda_0)]$.

Unfortunately, (10) does not have a closed-form solution for the group ridge estimator $[\hat{\boldsymbol{\beta}}_0(\lambda), \hat{\boldsymbol{\beta}}(\lambda)]$ in the general setting unless $b(\cdot)$ is in some specific form. The following corollary provides a closed-form solution for linear regression in which $b(\theta) = \theta^2/2$.

Corollary 5. *Consider linear regression in which $b(\theta) = \theta^2/2$ in (1). Further, suppose that both $\mathbf{Y}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are centralized so that the intercept β_0 is not included in (8). Then, (10) implies that*

$$(11)\qquad \hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda) = \left[\mathbf{I} - \frac{2(\lambda - \lambda_0)}{1 + 2\lambda\delta_\mathcal{A}} \mathcal{G}_\mathcal{A}^{-1} \mathbf{S}_\mathcal{A} \mathbf{S}_\mathcal{A}^\top\right] \hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_0),$$

where $\mathcal{G}_\mathcal{A} = \mathbb{X}_\mathcal{A}^\top \mathbb{X}_\mathcal{A}$ with $\mathbb{X}_\mathcal{A} = (\ldots, \mathbf{x}_k, \ldots)_{k \in \mathcal{A}}$, $\mathbf{S}_\mathcal{A} = (\ldots, S_k, \ldots)_{k \in \mathcal{A}}^\top$ with $S_k = \operatorname{sign}[\hat{\beta}_k(\lambda_0)]$, and $\delta_\mathcal{A} = \mathbf{S}_\mathcal{A}^\top \mathcal{G}_\mathcal{A}^{-1} \mathbf{S}_\mathcal{A}$.

Corollary 5 implies that the group ridge solution path follows the same "equiangular vector" direction, $\mathcal{G}_\mathcal{A}^{-1}\mathbf{S}_\mathcal{A}$, as in the lasso solution path [7]. However, it is piecewise smooth with respect to λ but not piecewise linear as in the lasso path. Although the two paths are different, the group ridge solution is still a monotone function of λ when $\mathcal{A}(\lambda)$ stays unchanged, the same as lasso. This can be verified by observing that its derivative with respect to λ has a fixed sign for each $k \in \mathcal{A}$:

$$\frac{d\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda)}{d\lambda} = -2\frac{1 + 2\lambda_0\delta_\mathcal{A}}{(1 + 2\lambda\delta_\mathcal{A})^2} \|\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_0)\|_1 \mathcal{G}_\mathcal{A}^{-1}\mathbf{S}_\mathcal{A}.$$

Next, let's discuss how the active set $\mathcal{A}(\lambda)$ changes between two consecutive steps. To facilitate the discussion, we will focus on the linear regression setting throughout the rest of this subsection, in which both $\mathbf{Y}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_d$ have been centralized and thus the intercept β_0 is not included in (8). Again, suppose the current step starts at λ_0 and $\mathcal{A}(\lambda_0) = \mathcal{A}$. There are two possible cases, i.e., an index k is either added to $\mathcal{A}$ or removed from $\mathcal{A}$. On the one hand, if $k \in \mathcal{A}$ is removed from $\mathcal{A}$ at λ, it is equivalent to $\hat{\beta}_k(\lambda) = 0$ since $\hat{\beta}_k(\lambda)$ is a monotone function of λ within the current step. In Corollary 5, we can solve the equation $\hat{\beta}_k(\lambda) = 0$ to get

$$\lambda_k^* = \frac{\hat{\beta}_k(\lambda_0) + 2\lambda_0 C_k \|\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_0)\|_1}{2C_k\|\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda_0)\|_1 - 2\hat{\beta}_k(\lambda_0)\delta_\mathcal{A}},$$

where $C_k = [\mathcal{G}_\mathcal{A}^{-1}\mathbf{S}_\mathcal{A}]_k$. Here, λ_k^* satisfies that $\hat{\beta}_k(\lambda_k^*) = 0$. Therefore, we define $\lambda_-^* = \max_{k \in \mathcal{A}}\{\lambda_k^* : \lambda_k^* \in (0, \lambda_0)\}$ to be the candidate λ where an index $k \in \mathcal{A}$ will be removed from $\mathcal{A}$.

On the other hand, we discuss when an index $k \notin \mathcal{A}$ will be added to $\mathcal{A}$. Let's define $a_k(\lambda) = \mathbf{x}_k^\top\{\mathbf{Y} - \mathbb{X}_\mathcal{A}\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda)\}$ for $\lambda \in (0, \lambda_0)$ and $b_k = \mathbf{x}_k^\top \mathbb{X}_\mathcal{A}\mathcal{G}_\mathcal{A}^{-1}\mathbf{S}_\mathcal{A}$. Furthermore, $\Delta_k(\lambda) = |a_k(\lambda)| - 2\lambda\|\hat{\boldsymbol{\beta}}_\mathcal{A}(\lambda)\|_1$ for $\lambda \in (0, \lambda_0)$. According to Theorem 1, $\Delta_k(\lambda) \le 0$ for $\lambda \in (0, \lambda_0)$ and all $k \notin \mathcal{A}$. For an index $k \notin \mathcal{A}$, if $\Delta_k(\lambda) = 0$ and $\Delta_k'(\lambda) < 0$, then k will be added to $\mathcal{A}$ at λ; otherwise, as λ further decreases to $\lambda-$, $\Delta_k(\lambda-) > 0$ and this violates the KKT conditions in Theorem 1. Therefore, to find out the value of λ at which an index $k \notin \mathcal{A}$ is added to $\mathcal{A}$, we just need to solve $\Delta_k(\lambda) = 0$ to get a candidate λ_k^* and verify that $\Delta_k'(\lambda_k^*) < 0$, where $\Delta_k'(\lambda)$ can be evaluated assisted by (11):

$$\Delta_k'(\lambda) = -2\frac{1+2\lambda_0\delta_{\mathcal{A}}}{(1+2\lambda\delta_{\mathcal{A}})^2}\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_0)\|_1\left(1 - b_k \operatorname{sign}[a_k(\lambda)]\right).$$

So, if $a_k(\lambda) > 0$, $\Delta_k'(\lambda) < 0$ is equivalent to $b_k < 1$ and $\Delta_k(\lambda) = 0$ leads to

$$\lambda_{k,1}^* = \frac{a_k(\lambda_0)/2 - \lambda_0 b_k\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_0)\|_1}{(1+2\lambda_0\delta_{\mathcal{A}} - b_k)\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_0)\|_1 - \delta_{\mathcal{A}}a_k(\lambda_0)}.$$

In parallel, if $a_k(\lambda) < 0$, $\Delta_k'(\lambda) < 0$ is equivalent to $b_k > -1$, and $\Delta_k(\lambda) = 0$ leads to

$$\lambda_{k,2}^* = \frac{-a_k(\lambda_0)/2 + \lambda_0 b_k\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_0)\|_1}{(1+2\lambda_0\delta_{\mathcal{A}} + b_k)\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\lambda_0)\|_1 + \delta_{\mathcal{A}}a_k(\lambda_0)}.$$

Thus, we define a candidate λ where an index $k \notin \mathcal{A}$ will be added to $\mathcal{A}$ as $\lambda_+^* = \max_{k\notin\mathcal{A}}[\{\lambda_{k,1}^* \in (0,\lambda_0) : b_k < 1\} \cup \{\lambda_{k,2}^* \in (0,\lambda_0) : b_k > -1\}]$.

Following the above discussion, a new step will start at $\lambda^* = \max(\lambda_-^*, \lambda_+^*)$, and an index will be removed from or added to $\mathcal{A}(\lambda^*)$ depending on whether λ_-^* or λ_+^* is the larger one, respectively. This result is summarized in Theorem 6. Together, Theorems 3–6 provide a thorough illustration of the group ridge solution path.

Theorem 6. *Consider linear regression in which $b(\theta) = \theta^2/2$ in (1). Further, suppose that both $\mathbf{Y}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are centralized so that the intercept β_0 is not included in (8). Then, there are two situations where $\mathcal{A}(\lambda)$ changes when λ decreases from λ_0. If $\lambda_-^* > \lambda_+^*$, an index $k \in \mathcal{A}$ is removed from $\mathcal{A}$ at $\lambda = \lambda_-^*$; otherwise, an index $k \notin \mathcal{A}$ is added to $\mathcal{A}$ at $\lambda = \lambda_+^*$.*

3. PRIOR GROUP BRIDGE

3.1 Definition

Real data often come with both prior-informative groups and groups without any prior information. As in the introduction, the house price study may not have other known factors besides school district and the genetic study may only have a small number of genes known to be associated with the disease. In these situations, we need to treat the prior-informative groups differently from the groups without prior information.

Let $\{A_j : j \in \mathcal{J}_1\}$ be the prior-informative groups while the others $\{A_j : j \in \mathcal{J}_2\}$ are not included in the prior knowledge. Here, $\mathcal{J}_1 \cup \mathcal{J}_2 = \{1, \ldots, J\}$ and $\mathcal{J}_1 \cap \mathcal{J}_2 = \emptyset$. Then it is natural to keep all the groups $\{A_j : j \in \mathcal{J}_1\}$ and to select important groups among $\{A_j : j \in \mathcal{J}_2\}$. To this end, we will impose the group ridge and group bridge penalties on prior-informative groups and groups without prior information, respectively. This leads to the following objective function:

$$(12) \qquad \ell(\beta_0, \boldsymbol{\beta}) = L(\beta_0, \boldsymbol{\beta}) + \lambda_1 \sum_{j\in\mathcal{J}_1} c_j\|\boldsymbol{\beta}_{A_j}\|_1^2 + \lambda_2 \sum_{j\in\mathcal{J}_2} c_j\|\boldsymbol{\beta}_{A_j}\|_1^\gamma,\ 0 < \gamma < 1,$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the tuning parameters and c_j's are constants to adjust the different dimensions of $\boldsymbol{\beta}_{A_j}$. Similar to group lasso and group bridge, we choose $c_j \propto |A_j|^{1/2}$ and $c_j \propto |A_j|^{1-\gamma}$ for $j \in \mathcal{J}_1$ and $j \in \mathcal{J}_2$, respectively. We refer to (12) as prior group bridge and call the minimizer of (12) the prior group bridge estimator.

Compared to group bridge, prior group bridge automatically includes the prior-informative groups in the model. Therefore, it is very crucial for us to ensure the prior information to be correct before it is incorporated into (12). Fortunately, such information is available in some real applications, such as replicable or well-known factors for a response in the two examples illustrated in the introduction.

3.2 Asymptotic properties

In Huang et al. [10], group bridge has been shown to possess the oracle property in terms of group selection. In other words, the group bridge estimators of the irrelevant groups are exactly equal to zero with probability converging to one and the group bridge estimators of the relevant groups are $\sqrt{n}$-consistent to their true coefficients.

As prior group bridge is a combination of group ridge and group bridge, it is not surprising to see the same oracle property holds for the groups without prior information, i.e., $\{A_j : j \in \mathcal{J}_2\}$. For prior-informative groups $\{A_j : j \in \mathcal{J}_1\}$, we need to assume the correctness of the prior information to ensure the selection consistency. This leads to the following main result for the asymptotic properties of prior group bridge.

Before presenting the main result, let us introduce some necessary notation as follows. Let $(\beta_{0,0}, \boldsymbol{\beta}_0)$ with an additional subscript 0 denote the true value of $(\beta_0, \boldsymbol{\beta})$ for clarity. Further, let B_2 be the union of the irrelevant groups with all zero coefficients and consequently $B_1 = \{1, \ldots, d\} \setminus B_2$ be the union of the relevant groups, each of which has at least one nonzero coefficient. Write $\boldsymbol{\beta}_{B_j} = (\beta_k : k \in B_j)^\top$ for $j = 1, 2$. Finally, define

$$\boldsymbol{\Sigma}(\beta_0, \boldsymbol{\beta}) = (\mathbf{1}, \mathbb{X})^\top \operatorname{diag}\{b''(\beta_0 + \mathbf{X}_1^\top\boldsymbol{\beta}), \ldots, b''(\beta_0 + \mathbf{X}_n^\top\boldsymbol{\beta})\}(\mathbf{1}, \mathbb{X}),$$

with $\mathbf{1} = (1, \ldots, 1)^\top$. As follows, Theorem 7 provides the asymptotic properties of the prior group bridge estimator.

Theorem 7. *Assume that $A_j \subseteq B_1$ for $j \in \mathcal{J}_1$. Suppose $n^{-1/2}\lambda_1 \to \lambda_1^* < \infty$, $n^{-1/2}\lambda_2 \to \lambda_2^* < \infty$,* *and* $n^{-\gamma/2}\lambda_2 \to \infty$. *Assume that* $\boldsymbol{\Sigma}(\beta_{0,0}, \boldsymbol{\beta}_0)/n \to \boldsymbol{\Sigma}^*$ *where* $\boldsymbol{\Sigma}^*$ *is a positive definite matrix and* $\sup\{\|\boldsymbol{\Sigma}(b_0, \mathbf{b}) - \boldsymbol{\Sigma}(\beta_{0,0}, \boldsymbol{\beta}_0)\|_2 : \sqrt{n}\|\{b_0 - \beta_{0,0}, (\mathbf{b} - \boldsymbol{\beta}_0)^\top\}\|_2 \le \delta\} \to 0$ *for any* $\delta > 0$. *Then,*

(a) $P(\hat{\boldsymbol{\beta}}_{B_2} = \mathbf{0}) \to 1$;

(b) $[\sqrt{n}(\hat{\beta}_0 - \boldsymbol{\beta}_{0,0}), \sqrt{n}(\hat{\boldsymbol{\beta}}_{B_1} - \boldsymbol{\beta}_{B_1,0})] \xrightarrow{d} \arg\min_{u_0,\mathbf{u}} V(u_0, \mathbf{u})$, *where*

$$V(u_0, \mathbf{u}) = -(u_0, \mathbf{u}^\top)\mathbf{Z} + \frac{1}{2}(u_0, \mathbf{u}^\top)\boldsymbol{\Sigma}^*_{1,1}(u_0, \mathbf{u}^\top)^\top$$
$$+ 2\lambda_1^* \sum_{j\in\mathcal{J}_1} c_j \|\boldsymbol{\beta}_{A_j,0}\|_1 \times$$
$$\sum_{k\in A_j\cap B_1} \{u_k \operatorname{sign}(\beta_{k,0}) I(\beta_{k,0} \ne 0) + |u_k| I(\beta_{k,0} = 0)\}$$
$$+ \gamma\lambda_2^* \sum_{j\in\mathcal{J}_2} c_j \|\boldsymbol{\beta}_{A_j,0}\|_1^{\gamma-1} \times$$
$$\sum_{k\in A_j\cap B_1} \{u_k \operatorname{sign}(\beta_{k,0}) I(\beta_{k,0} \ne 0) + |u_k| I(\beta_{k,0} = 0)\}$$

with $\mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^*_{1,1})$ *and* $\boldsymbol{\Sigma}^*_{1,1}$ *being the* $(\{0,B_1\},\{0,B_1\})$-*submatrix of* $\boldsymbol{\Sigma}^*$. *In particular, when* $\lambda_1^* = \lambda_2^* = 0$,

$$[\sqrt{n}(\hat{\beta}_0 - \boldsymbol{\beta}_{0,0}), \sqrt{n}(\hat{\boldsymbol{\beta}}_{B_1} - \boldsymbol{\beta}_{B_1,0})] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^{*-1}_{1,1}).$$

The assumption $A_j \subseteq B_1$ for $j \in \mathcal{J}_1$ reflects our requirement about the prior information. The prior-informative groups cannot be irrelevant to the response to avoid incorrect group inclusion. Under this assumption, prior group bridge also possesses the oracle property in terms of group selection just like group bridge. The major difference between prior group bridge and group bridge lies in the asymptotic distribution of the estimators $\{\hat{\beta}_{A_j} : j \in \mathcal{J}_1\}$ as the penalties imposed on them are different. In the special case when $\lambda_1^* = \lambda_2^* = 0$, the two estimators have the same asymptotic distribution.

Although both methods have the same and optimal theoretical property, i.e., the oracle property, group bridge can still miss a relevant group with small coefficients in practice (see Section 4). Weak signals are actually common in many studies. For example, genetic effects tend to have a small size and are notoriously hard to identify in a single genome-wide association study [14]. Fortunately, prior genetic research has accumulated a large amount of valuable knowledge, e.g., in GWAS Catalog [26]. To avoid potential loss of signal identifications, prior group bridge is preferred over group bridge as it can incorporate the valuable prior information and help identify the weak genetic effects. We will illustrate it further in our simulation studies and real data analysis.

3.3 Computational algorithms

To minimize (12), we optimize the objective function iteratively with respect to one of the following parameter sets, β_0, $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_1\}$, and $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_2\}$, regarding the other two parameter sets as fixed. To update β_0, we simply take the derivative of $L(\beta_0, \boldsymbol{\beta})$ with respect to β_0 and set it to be zero, because both penalties do not involve β_0. To update $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_1\}$, the optimization can be regarded as a group ridge problem as the objective function only involves a group ridge penalty. Similarly, the optimization with respect to $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_2\}$ can be regarded as a group bridge problem as it only involves a group bridge penalty.

For group ridge, we derive a coordinate descent algorithm from the KKT conditions of (12) similar to those in Theorem 1. The coordinate descent algorithm updates one parameter at a time by regarding all other parameters as fixed. For each coordinate k, the algorithm either sets β_k to be 0 or updates β_k by solving an equation from the KKT conditions depending on whether the condition for $\beta_k = 0$ is satisfied or not. For group bridge, we follow the algorithm in Huang et al. [10] as well as the optimization function `gBridge` in the R package `grpreg`. It is noteworthy that both optimizations for $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_1\}$ and $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_2\}$ have an offset that needs to be adjusted in prior group bridge. The offset when optimizing $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_1\}$ is $\beta_0\mathbf{1} + \sum_{j\in\mathcal{J}_2} \mathbb{X}_{A_j}\boldsymbol{\beta}_{A_j}$ and the offset when optimizing $\{\boldsymbol{\beta}_{A_j} : j \in \mathcal{J}_2\}$ is $\beta_0\mathbf{1} + \sum_{j\in\mathcal{J}_1} \mathbb{X}_{A_j}\boldsymbol{\beta}_{A_j}$.

4. SIMULATION STUDIES

We use simulation studies to evaluate the performance of prior group bridge and compare it with other methods such as group bridge and lasso.

We adopt two simulation settings: (a) “small n small p” in which the sample size $n = 100$ for linear regression, $n = 200$ for logistic regression, and the number of parameters $p = 36$; (b) “large n large p” in which the sample size $n = 500$ for linear regression, $n = 1{,}000$ for logistic regression, and the number of parameters $p = 180$. In each setting, for $i = 1, \ldots, n$, we first simulate a covariate vector $\mathbf{X}$ of length p from multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with covariances $\sigma_{i,j} = \rho^{|i-j|}$ for $0 < \rho < 1$ and $i, j = 1, \ldots, p$. To impose a group structure on the covariates, write $\mathbf{X} = (\mathbf{X}_{A_1}^\top, \mathbf{X}_{A_2}^\top, \ldots, \mathbf{X}_{A_6}^\top)^\top$ into six groups with alternating sizes 8 and 4 in the “small n small p” setting, and with alternating sizes 40 and 20 in the “large n large p” setting. Based on the simulated covariate vector $\mathbf{X}$, we further simulate a response Y using either a linear model or a logistic regression model:

$$Y \sim N(\mu, 1), \quad \text{with } \mu = \mathbf{X}^\top\boldsymbol{\beta};$$
$$\text{or } Y \sim \text{Bernoulli}(\mu), \quad \text{with } \operatorname{logit}(\mu) = \mathbf{X}^\top\boldsymbol{\beta}.$$

In the “small n small p” setting, we set coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_{A_1}^\top, \boldsymbol{\beta}_{A_2}^\top, \ldots, \boldsymbol{\beta}_{A_6}^\top)^\top$, corresponding to the six groups of $\mathbf{X}$, as in Table 1. In the “large n large p” setting, we keep the group number as six while replicating the coefficients in each group in the “small n small p” setting five times. It is seen that the first four groups are relevant to the response while the last two are irrelevant. In addition, groups A_1 and A_2 possess sparse and relatively strong signals and groups A_3 and A_4 possess sparse and relatively weak signals.

In this simulation study, we compare three grouped variable selection methods—prior group bridge with complete prior information (i.e., $\mathcal{J}_1 = \{1, 2, 3, 4\}$), prior group bridge with incomplete prior information (i.e., $\mathcal{J}_1 = \{1, 4\}$), group bridge where γ is set to be 0.5 in the group bridge penalty, and a variable selection method, lasso. For all the methods, we use the Bayesian Information Criterion (BIC) to determine their optimal tuning parameters. We run the simulation for 5,000 times and summarize the following results:

(a) TPG, FPG: the number of true-positive and false-positive groups;
(b) TPV, FPV: the number of true-positive and false-positive variables;
(c) SE: the estimation error measured by the squared error loss;
(d) PE: the prediction error evaluated on the test data, measured by the mean of squared residuals for linear regression and misclassification error rate for logistic regression.

Table 1. Coefficients in the "small n small p" setting

Group	Size	Coefficients
A_1	8	$\boldsymbol{\beta}_{A_1} = (1.5, 3, 0, \ldots, 0)^\top$
A_2	4	$\boldsymbol{\beta}_{A_2} = (1, 2, 0, 0)^\top$
A_3	8	$\boldsymbol{\beta}_{A_3} = \boldsymbol{\beta}_{A_1}/5$
A_4	4	$\boldsymbol{\beta}_{A_4} = \boldsymbol{\beta}_{A_2}/5$
A_5	8	$\boldsymbol{\beta}_{A_5} = (0, 0, \ldots, 0)^\top$
A_6	4	$\boldsymbol{\beta}_{A_6} = (0, 0, 0, 0)^\top$

Tables 2 and 3 summarize the simulation results for linear regression and logistic regression in the "small n small p" setting, respectively; Tables 4 and 5 summarize the simulation results for linear regression and logistic regression in the "large n large p" setting, respectively. For linear regression (Tables 2 and 4), it is seen that prior group bridge outperforms group bridge in almost all aspects. Compared to group bridge, prior group bridge recovers more relevant groups and variables, includes fewer irrelevant variables, and results in more accurate estimation and prediction. Both methods perform very well in terms of excluding irrelevant groups from the model. In addition, comparing prior group bridge with complete and incomplete prior information, the extra information helps prior group bridge achieve better performance in terms of group and variable selection, estimation, and prediction as expected. Comparing lasso with grouped variable selection methods, lasso keeps almost all groups including the irrelevant ones. This is because lasso does not perform any group-level selection. Lasso also selects many more irrelevant variables than the three grouped variable selection methods, leading to a slightly worse performance in estimation and prediction.

The results for logistic regression (Tables 3 and 5) have a similar trend to those for linear regression. Prior group bridge clearly outperforms group bridge in terms of group selection because group bridge totally ignores the two groups that possess weak signals (A_3 and A_4). For variable selection, prior group bridge recovers more relevant variables in all settings but also keeps slightly more irrelevant variables. Both methods result in similar estimation accuracies relative to their large standard errors and prior group bridge gives slightly smaller misclassification errors when applied to test data. This is probably because the magnitude of the coefficients in groups A_3 and A_4 is small even

Table 2. Simulation results for linear regression in the "small n small p" setting. Reported are averages from 5,000 replications with standard errors in the parentheses. pgbridge.c: prior group bridge with complete prior information; pgbridge.i: prior group bridge with incomplete prior information; gbridge: group bridge

Method	TPG	FPG	TPV	FPV	SE	PE
$\rho = 0.1$						
pgbridge.c	4.0 (0.0)	1.1e-1 (4.7e-3)	7.5 (9.5e-3)	2.0 (2.4e-2)	0.19 (1.6e-3)	1.2 (2.9e-3)
pgbridge.i	3.9 (4.1e-3)	5.0e-2 (3.2e-3)	7.4 (1.1e-2)	2.9 (2.2e-2)	0.24 (2.3e-3)	1.3 (3.6e-3)
gbridge	3.7 (7.9e-3)	2.0e-2 (1.8e-3)	6.9 (1.6e-2)	4.7 (3.1e-2)	0.32 (2.3e-3)	1.3 (3.7e-3)
lasso	4.0 (6.3e-4)	1.7 (7.0e-3)	7.8 (6.4e-3)	10 (6.5e-2)	0.25 (1.5e-3)	1.3 (3.0e-3)
$\rho = 0.5$						
pgbridge.c	4.0 (0.0)	1.1e-1 (4.6e-3)	7.4 (9.6e-3)	1.9 (2.2e-2)	0.24 (2.2e-3)	1.2 (2.8e-3)
pgbridge.i	3.9 (3.9e-3)	4.0e-2 (3.1e-3)	7.2 (1.1e-2)	2.5 (2.1e-2)	0.31 (3.0e-3)	1.2 (3.6e-3)
gbridge	3.9 (5.3e-3)	2.0e-2 (1.8e-3)	7.1 (1.3e-2)	4.2 (2.7e-2)	0.34 (2.5e-3)	1.3 (3.3e-3)
lasso	4.0 (2.0e-4)	1.6 (8.2e-3)	7.8 (6.5e-3)	8.6 (6.0e-2)	0.26 (1.9e-3)	1.2 (2.8e-3)
$\rho = 0.9$						
pgbridge.c	4.0 (0.0)	1.1e-1 (4.7e-3)	6.9 (1.1e-2)	1.1 (1.8e-2)	0.96 (1.1e-2)	1.2 (2.7e-3)
pgbridge.i	3.8 (5.1e-3)	1.0e-1 (4.4e-3)	6.3 (1.3e-2)	2.7 (2.1e-2)	1.2 (1.1e-2)	1.2 (2.9e-3)
gbridge	3.5 (7.3e-3)	5.0e-2 (3.2e-3)	6.1 (1.4e-2)	4.7 (2.7e-2)	1.4 (1.1e-2)	1.3 (3.0e-3)
lasso	4.0 (1.2e-3)	1.3 (1.1e-2)	7.1 (1.0e-2)	7.1 (5.1e-2)	0.90 (8.6e-3)	1.2 (2.6e-3)

Table 3. Simulation results for logistic regression in the "small n small p" setting. Reported are averages from 5,000 replications with standard errors in the parentheses. pgbridge.c: prior group bridge with complete prior information; pgbridge.i: prior group bridge with incomplete prior information; gbridge: group bridge

Method	TPG	FPG	TPV	FPV	SE	PE (%)
$\rho = 0.1$						
pgbridge.c	4.0 (0.0)	7.0e-2 (5.0e-3)	6.4 (2.0e-2)	2.4 (2.6e-2)	2.5 (4.6e-2)	14 (5.4e-2)
pgbridge.i	3.2 (7.3e-3)	2.8e-2 (3.2e-3)	5.5 (1.7e-2)	1.4 (2.3e-2)	2.0 (3.6e-2)	15 (5.0e-2)
gbridge	2.1 (6.3e-3)	4.3e-3 (1.3e-3)	4.1 (1.0e-2)	0.65 (2.1e-2)	1.7 (2.8e-2)	15 (5.0e-2)
lasso	3.9 (4.6e-3)	1.8 (8.4e-3)	6.8 (1.7e-2)	11 (7.5e-2)	2.6 (2.0e-2)	15 (5.1e-2)
$\rho = 0.5$						
pgbridge.c	4.0 (0.0)	6.2e-2 (4.7e-3)	6.0 (2.1e-2)	1.6 (2.4e-2)	3.0 (4.9e-2)	13 (5.2e-2)
pgbridge.i	3.2 (7.7e-3)	2.7e-2 (3.1e-3)	5.2 (1.9e-2)	1.3 (2.2e-2)	2.6 (4.1e-2)	13 (5.0e-2)
gbridge	2.2 (8.0e-3)	6.2e-3 (1.5e-3)	4.1 (1.4e-2)	0.79 (2.4e-2)	2.2 (3.3e-2)	13 (4.7e-2)
lasso	3.9 (5.1e-3)	1.7 (9.2e-3)	6.7 (1.7e-2)	9.4 (6.7e-2)	2.7 (2.0e-2)	13 (4.8e-2)
$\rho = 0.9$						
pgbridge.c	4.0 (0.0)	1.2e-2 (2.2e-3)	5.2 (1.9e-2)	0.75 (1.6e-2)	7.6 (7.1e-2)	9.1 (4.1e-2)
pgbridge.i	3.0 (3.7e-3)	1.1e-2 (2.0e-3)	4.3 (1.6e-2)	1.0 (1.7e-2)	6.6 (6.9e-2)	9.2 (4.1e-2)
gbridge	2.2 (8.0e-3)	1.5e-2 (2.3e-3)	3.6 (1.5e-2)	1.8 (2.4e-2)	6.1 (6.1e-2)	9.5 (4.2e-2)
lasso	3.7 (8.5e-3)	1.4 (1.2e-2)	5.4 (1.9e-2)	6.2 (4.6e-2)	4.7 (3.7e-2)	9.1 (4.1e-2)

Table 4. Simulation results for linear regression in the "large n large p" setting. Reported are averages from 5,000 replications with standard errors in the parentheses. pgbridge.c: prior group bridge with complete prior information; pgbridge.i: prior group bridge with incomplete prior information; gbridge: group bridge

Method	TPG	FPG	TPV	FPV	SE	PE
$\rho = 0.1$						
pgbridge.c	4.0 (0.0)	2.0e-3 (6.0e-4)	40.0 (2.2e-3)	6.6 (3.4e-2)	5.4e-2 (2.0e-4)	1.0 (1.0e-3)
pgbridge.i	4.0 (0.0)	0.0 (0.0)	39.9 (1.0e-3)	7.0 (2.8e-2)	5.6e-2 (2.0e-4)	1.0 (1.0e-3)
gbridge	4.0 (0.0)	0.0 (0.0)	39.9 (5.5e-3)	6.8 (4.4e-2)	7.2e-2 (3.0e-4)	1.1 (1.0e-3)
lasso	4.0 (0.0)	2.0 (2.0e-4)	40.0 (0.0)	52 (1.5e-1)	1.1e-1 (3.0e-4)	1.1 (1.0e-3)
$\rho = 0.5$						
pgbridge.c	4.0 (0.0)	1.0e-3 (5.0e-4)	40.0 (4.0e-3)	5.6 (3.2e-2)	7.0e-2 (3.0e-4)	1.1 (1.0e-3)
pgbridge.i	4.0 (0.0)	0.0 (0.0)	39.9 (2.1e-3)	6.5 (2.9e-2)	8.0e-2 (3.0e-4)	1.1 (1.0e-3)
gbridge	4.0 (0.0)	0.0 (0.0)	39.9 (4.4e-3)	6.1 (3.4e-2)	8.0e-2 (4.0e-4)	1.1 (1.0e-3)
lasso	4.0 (0.0)	2.0 (8.0e-4)	40.0 (0.0)	43 (1.4e-1)	1.1e-1 (3.0e-4)	1.1 (1.0e-3)
$\rho = 0.9$						
pgbridge.c	4.0 (0.0)	1.0e-3 (6.0e-4)	38.3 (2.1e-2)	12 (4.4e-2)	0.54 (3.2e-3)	1.1 (1.0e-3)
pgbridge.i	4.0 (0.0)	0.0 (0.0)	38.0 (1.9e-2)	13 (4.4e-2)	0.59 (2.5e-3)	1.1 (1.0e-3)
gbridge	4.0 (0.0)	0.0 (0.0)	38.8 (9.3e-3)	24 (6.2e-2)	0.62 (1.9e-3)	1.1 (1.0e-3)
lasso	4.0 (0.0)	1.4 (9.4e-3)	39.9 (5.4e-3)	26 (6.6e-2)	0.45 (1.5e-3)	1.1 (1.0e-3)

though group bridge totally ignores them. The performance of lasso is similar to that for linear regression. It underperforms all three grouped variable selection methods with a much higher number of irrelevant groups/variables and slightly lower accuracy in terms of estimation and prediction.

In addition, we examine the performance of all methods in terms of variable selection within each group. Figures 1 and 2 present the boxplots of the number of true and false positives within all groups, $A_1, \ldots, A_6$, for linear regression and logistic regression, respectively, in the "small n small p" setting; Figures 3 and 4 present the same results in the "large n large p" setting. These results further confirm the findings from Tables 2–5. For linear regression, group bridge tends to have more false positives especially in groups with strong signals (A_1 and A_2). For logistic regression, group bridge totally ignores the groups with weak signals (A_3 and A_4). Lasso always selects more irrelevant groups and variables than the three grouped variable selection methods. It is clearly seen that prior group bridge performs better in terms of variable selection with the assistance from the prior knowledge.

It is noteworthy that the above conclusions drawn from comparing different methods are consistent in both the "small n small p" and the "large n large p" settings, suggesting that the advantage of prior group bridge over its competitors is robust to both the size and the dimension of the data.

Table 5. Simulation results for logistic regression in the "large n large p" setting. Reported are averages from 5,000 replications with standard errors in the parentheses. pgbridge.c: prior group bridge with complete prior information; pgbridge.i: prior group bridge with incomplete prior information; gbridge: group bridge

Method	TPG	FPG	TPV	FPV	SE	PE (%)
$\rho = 0.1$						
pgbridge.c	4.0 (0.0)	0.0 (0.0)	26.6 (1.7e-1)	10 (1.4e-1)	25 (0.59)	13 (7.9e-2)
pgbridge.i	3.0 (0.0)	0.0 (0.0)	24.1 (8.6e-2)	6.2 (7.8e-2)	13 (0.36)	11 (5.6e-2)
gbridge	2.0 (0.0)	0.0 (0.0)	18.8 (3.9e-2)	1.5 (6.5e-2)	11 (0.15)	11 (4.2e-2)
lasso	4.0 (0.0)	2.0 (0.0)	33.4 (7.3e-2)	56 (3.0e-1)	30 (0.19)	11 (4.4e-2)
$\rho = 0.5$						
pgbridge.c	4.0 (0.0)	0.0 (0.0)	23.0 (1.4e-1)	7.5 (1.1e-1)	24 (0.48)	11 (6.3e-2)
pgbridge.i	3.0 (0.0)	0.0 (0.0)	21.8 (1.1e-1)	5.0 (7.7e-2)	16 (0.27)	11 (5.3e-2)
gbridge	2.0 (0.0)	0.0 (0.0)	18.9 (3.5e-2)	2.9 (5.8e-2)	13 (0.15)	9.7 (3.9e-2)
lasso	4.0 (0.0)	2.0 (0.0)	32.7 (7.3e-2)	48 (2.7e-1)	30 (0.18)	9.3 (3.7e-2)
$\rho = 0.9$						
pgbridge.c	4.0 (0.0)	0.0 (0.0)	14.3 (9.1e-2)	7.8 (9.2e-2)	62 (0.42)	7.8 (6.1e-2)
pgbridge.i	3.0 (1.2e-3)	0.0 (0.0)	13.8 (8.4e-2)	7.1 (8.5e-2)	53 (0.43)	7.5 (4.3e-2)
gbridge	2.1 (8.3e-3)	0.0 (0.0)	13.1 (8.4e-2)	7.5 (9.1e-2)	46 (0.36)	8.0 (3.9e-2)
lasso	4.0 (0.0)	2.0 (3.2e-3)	23.7 (8.6e-2)	30 (1.6e-1)	42 (0.16)	5.4 (2.8e-2)

In summary, equipped with the prior information, prior group bridge possesses a better overall performance than the method without incorporating such information. For those prior-informative groups, the group ridge penalty is obviously a more natural choice than group bridge as it eliminates the possibility of losing true signals. More interestingly, prior group bridge tends to perform better in terms of variable selection within groups as well, as shown in Figures 1–4. Incorporating correct information also slightly increases the accuracies of parameter estimation and prediction.

5. REAL DATA

In this section, we apply prior group bridge to a real dataset from a genetic study of bipolar disorder. Bipolar disorder is a serious and potentially life-threatening mood disorder [16] and it is well known that bipolar disorder has a substantial genetic component [6]. There have been at least six genome-wide association studies reported in the literature so far [5, 1, 8, 21, 19, 22]. While these studies revealed promising association signals, top findings from various studies do not show obvious overlap [2, 17].

In this work, we analyze the SNP data collected by the Wellcome Trust Case Control Consortium (WTCCC) as the primary dataset [5]. To collect prior information, we choose the genetic findings from the five other studies. Among the five studies, Ferreira et al. [8] and Scott et al. [19] used meta-analysis to jointly analyze the WTCCC data and another dataset; Smith et al. [22] analyzed data from two populations, i.e., individuals of European ancestry and African ancestry. Since the WTCCC dataset is our primary dataset and it only includes individuals of European ancestry, we omit the results from the above three studies. Therefore, we summarize the results from Baum et al. [1] and Sklar et al. [21] as prior genetic risk factors for bipolar disorder.

We incorporate the prior genetic findings at the gene level instead of the SNP level for two reasons. First, different genotyping platforms were used in these two prior studies. Baum et al. [1] used the Illumina HumanHap550 Array and Sklar et al. [21] used the Affymetrix Human Mapping 500K Array. This results in many non-overlapping SNPs between the two datasets. Second, it is more reasonable to assume that gene-level signals are preserved among different studies than SNP-level signals. A recent large-scale GWAS meta-analysis [18] found that many of the risk genes are shared by different studies, but the SNP-level association signals can be distinct. Due to the above reasons, we summarize the prior genetic factors into risk genes instead of risk SNPs.

Recall that we need to ensure the correctness of the prior information before incorporating it into prior group bridge for good performance. To this end, we only select the very top findings that have been discovered as well as validated by replication samples. After a careful selection, only seven risk genes fall into this category: DFNB31, DGKH, EGFR, MYO5B, NALCN, NXN, and SORCS2. We believe that these seven genetic signals are qualified to serve as prior information and thus regard them as the prior-informative genes (groups) for the WTCCC SNP data.

In addition to collecting prior information, we prescreen the WTCCC data before carrying out the association study. As in most genetic association studies, all SNPs with a minor allele frequency less than 0.05 or failing the Hardy-Weinberg equilibrium test at p-value 0.0001 are excluded from further analysis. To further reduce the computational burden, a univariate SNP analysis was performed to select a smaller set of SNPs. Specifically, a logistic regression is run separately for each SNP and a SNP is included in the final analysis

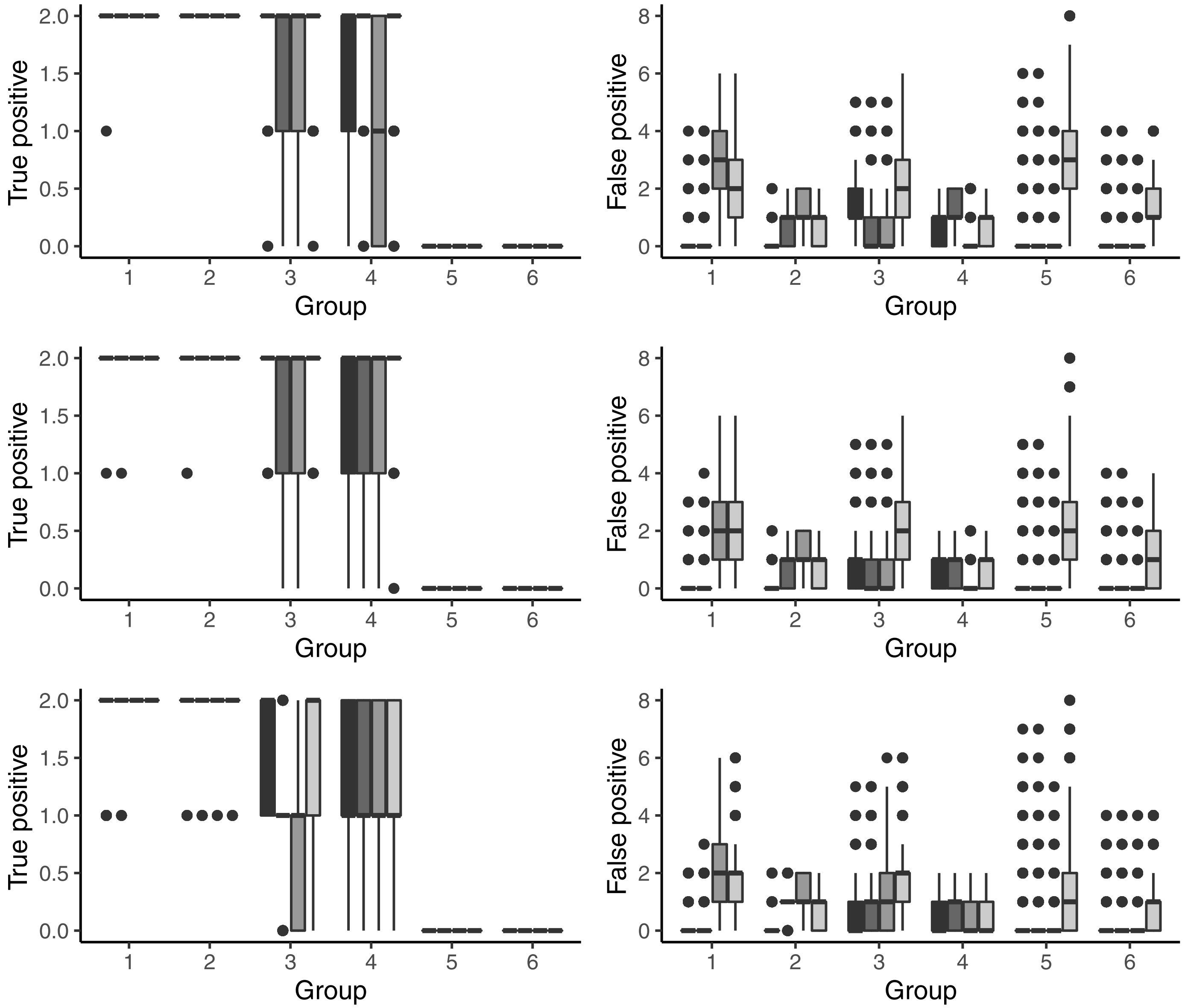

Figure 1. Number of true positives and false positives within each group for linear regression in the "small n small p" setting, selected by prior group bridge with complete prior information (dark), prior group bridge with incomplete prior information (dark grey), group bridge (light grey), and lasso (light). The three rows correspond to $\rho = 0.1$, 0.5, and 0.9, respectively.

if either of the following two conditions is satisfied: (a) its p-value from the logistic regression is smaller than 0.05; (b) it belongs to one of the seven prior-informative genes and its logistic regression p-value is smaller than 0.1.

After prescreening, we use SNPnexus (http://snpnexus.org), an online genomic mapping tool, to map the SNPs in the WTCCC data to their genes including the seven risk genes. This mapping procedure results in our final dataset with 10717 SNPs mapped to 3582 genes, among which 25 SNPs belong to the seven prior-informative genes. Moreover, the final dataset includes 5002 samples with 1998 cases and 3004 controls. We then apply both prior group bridge and group bridge to this dataset with their tuning parameters selected by BIC as in the simulation studies.

Table 6 summarizes our findings within the seven prior-informative genes. The advantage of prior group bridge over group bridge is obvious: prior group bridge is able to keep these important genes by identifying risk SNPs within them, while group bridge totally ignores all these well-known genetic signals. In addition, the estimated odds ratios from a refitted model imply that most of the SNP-level signals in these risk genes are relatively weak. This is probably why group bridge ignores all these risk genes, similar to what we have observed in simulations. In other words, group bridge

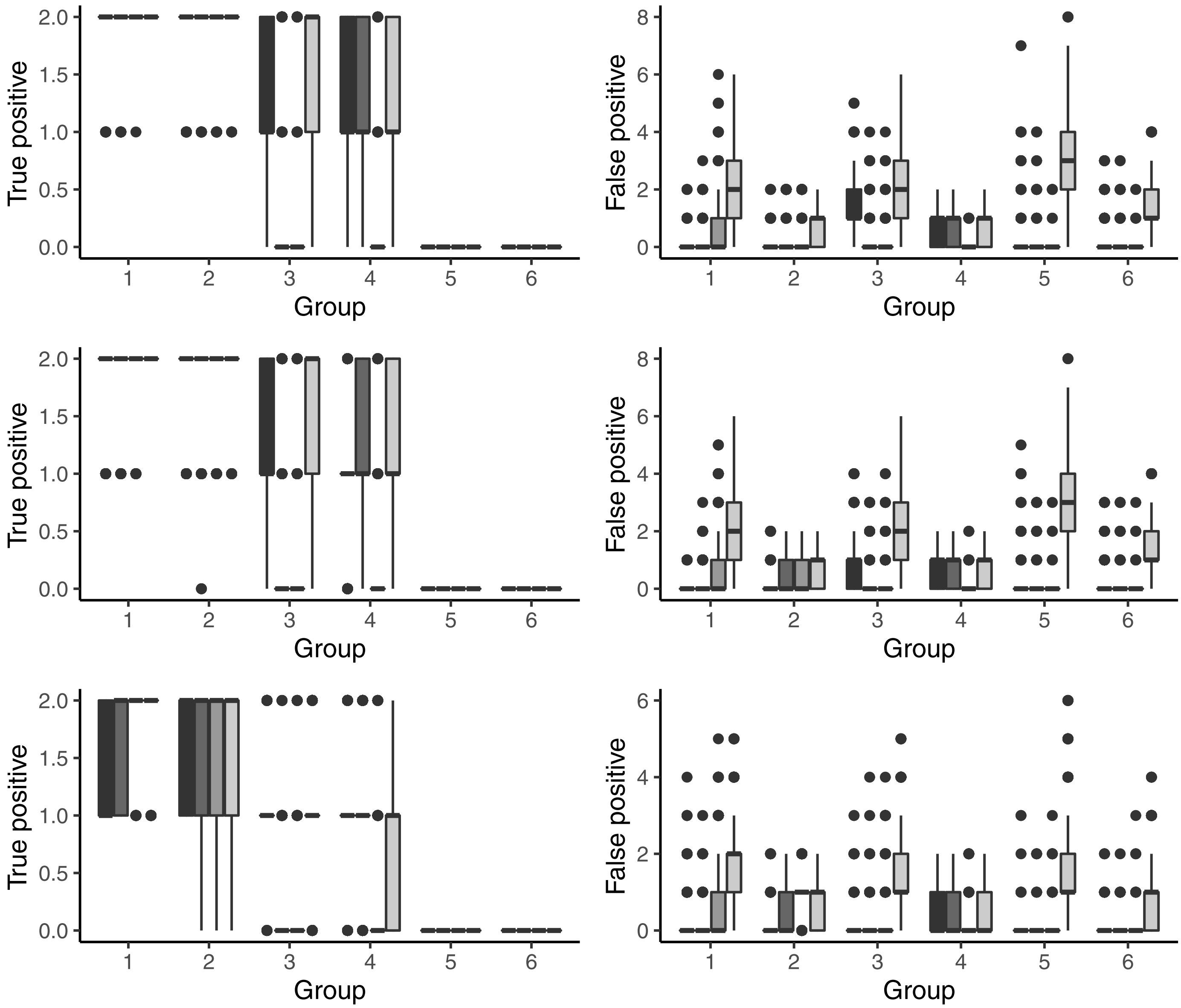

Figure 2. Number of true positives and false positives within each group for logistic regression in the "small n small p" setting, selected by prior group bridge with complete prior information (dark), prior group bridge with incomplete prior information (dark grey), and group bridge (light grey), and lasso (light). The three rows correspond to $\rho = 0.1$, 0.5, and 0.9, respectively.

tends to ignore a relevant group with small individual coefficients.

For the genes without prior information, we summarize our findings at the SNP level and at the gene level separately. At the SNP level, both prior group bridge and group bridge select a large number of SNPs, with 39 SNPs by prior group bridge and 40 SNPs by group bridge. Interestingly, 25 SNPs are overlapped between those selected by the two methods, while prior group bridge selects 14 additional SNPs and group bridge selects 15 additional SNPs. At the gene level, prior group bridge and group bridge identify 25 common genes, while prior group bridge identifies 13 additional genes and group bridge identifies 15 additional genes. In summary, there are a lot of similarities between prior group bridge and group bridge in terms of selection of SNPs and genes without any prior information.

In summary, the genetic factors identified by prior group bridge are consistent with the known fact that the seven risk genes are associated with bipolar disorder, but group bridge results in inconsistent findings. Additionally, the two methods perform fairly similarly for the genes with no prior information. Finally, the large number of genetic signals identified by both methods suggests that bipolar disorder has a very complex genetic etiology. It is not a single gene or a

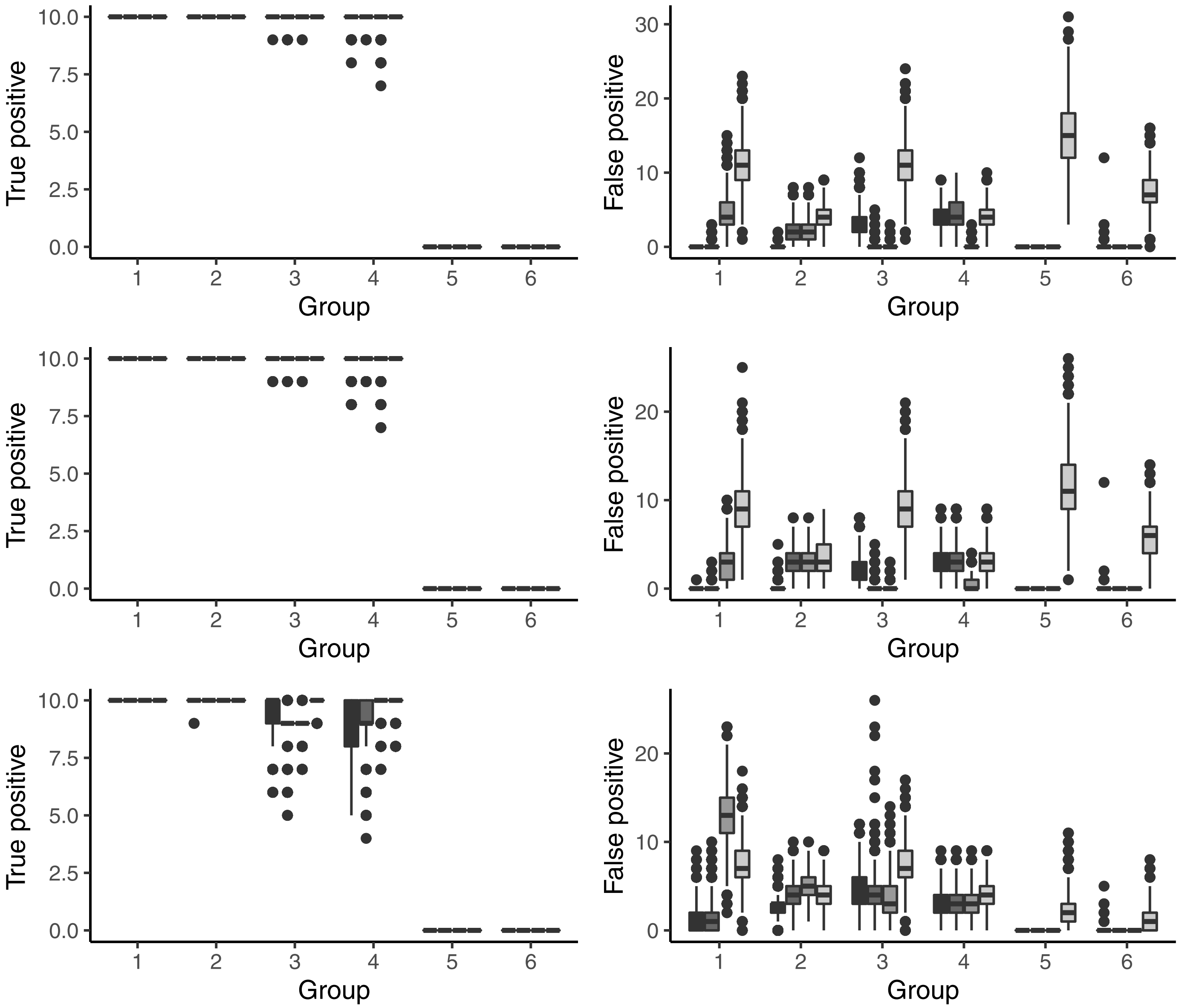

Figure 3. Number of true positives and false positives within each group for linear regression in the "large n large p" setting, selected by prior group bridge with complete prior information (dark), prior group bridge with incomplete prior information (dark grey), group bridge (light grey), and lasso (light). The three rows correspond to $\rho = 0.1$, 0.5, and 0.9, respectively.

small number of genes that contribute to the disease risk. It might involve many risk genes, each of which only contributes a moderate or even a weak effect.

6. DISCUSSION

In this work, we propose a group penalty called group ridge to select at least one variable in a prior-informative group, different from other group penalties. Furthermore, we develop prior group bridge by applying group ridge and group bridge to groups with and without prior information in a multiple regression problem, respectively. Our study shows that the incorporation of correct prior information improves the performance in group and variable selections. Other group penalizations, such as group bridge, can omit a whole group, even if it is a relevant group with nonzero coefficients.

Due to the nature of group ridge, we need to ensure the correctness of prior information before applying prior group bridge to real data. Otherwise, the wrong group-level signals from the prior information will be included in the model. In some real problems, this information is available from either the experts or studies that have been performed previously as illustrated in both introduction and real data sections.

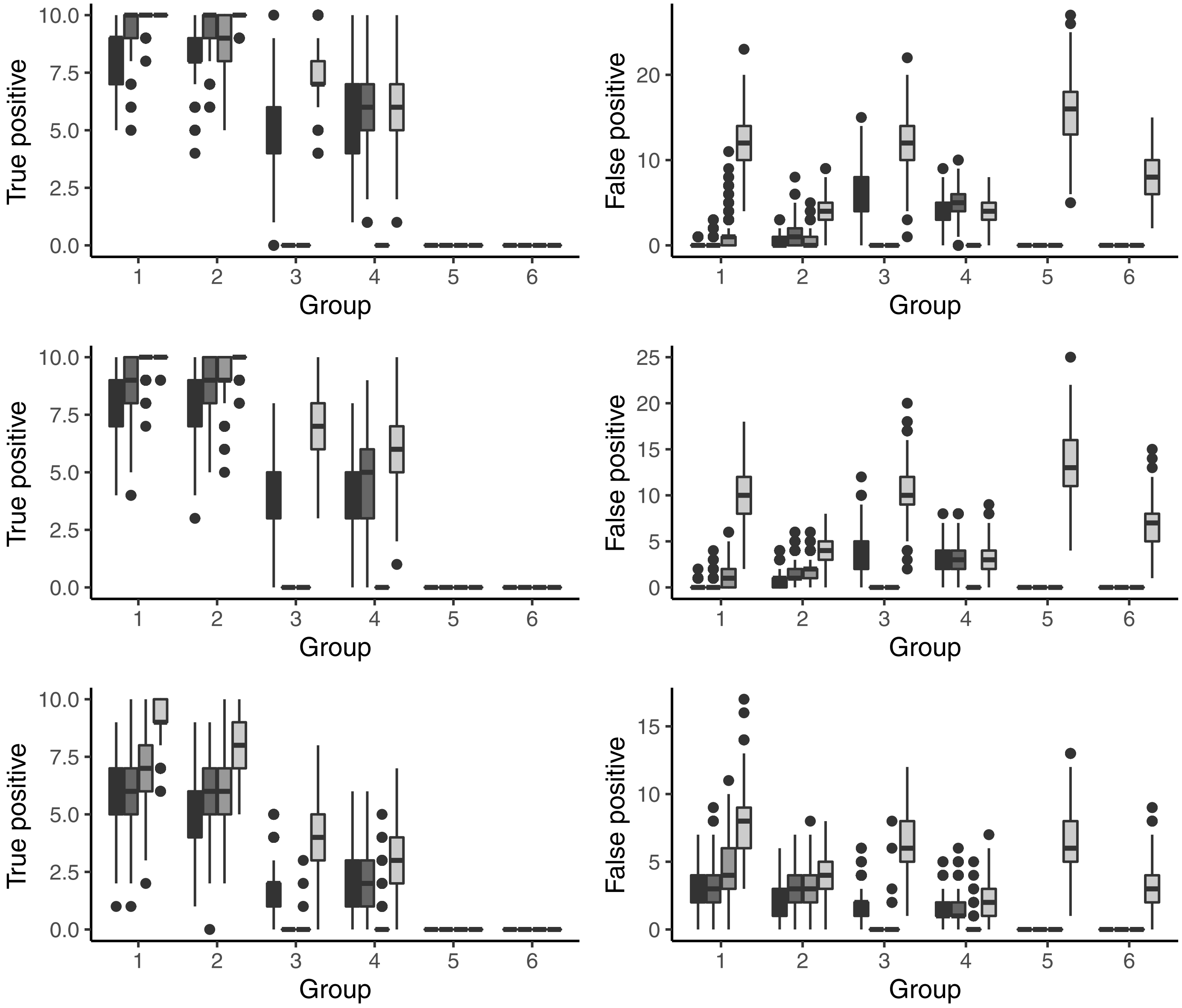

Figure 4. Number of true positives and false positives within each group for logistic regression in the "large n large p" setting, selected by prior group bridge with complete prior information (dark), prior group bridge with incomplete prior information (dark grey), and group bridge (light grey), and lasso (light). The three rows correspond to $\rho = 0.1$, 0.5, and 0.9, respectively.

Ignorance of such valuable information will result in potentially inconsistent findings and possible loss of estimation and prediction accuracies.

A question naturally arises following the current work: how can one extend the proposed approach to allow possibly wrong information? One possible way is to follow the idea in Jiang, He and Zhang [11], in which the authors proposed a two-step approach to incorporating variable-level prior information with various qualities. They first incorporate the information as it is correct and then balance between the prior information and the data. Similarly, we can add a balancing step after we apply prior group bridge. However, this is beyond the scope of the current work and we will leave it to future investigations.

APPENDIX

Proof of Theorem 3

From the objective function (8) and its associated KKT conditions, $\hat{\boldsymbol{\beta}}(\lambda) \neq \mathbf{0}$ for any $\lambda > 0$. Otherwise, according to (6), $\sum_{i=1}^{n}[Y_i - b'(\hat{\beta}_0(\lambda) + \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}(\lambda))]X_{ik} = 0$ for all $k = 1, \ldots, d$, which in conjunction with (5) implies $\hat{\boldsymbol{\beta}}(\lambda)$ being the unpenalized estimator and a contradiction to $\hat{\boldsymbol{\beta}}(\lambda) = \mathbf{0}$.

Table 6. Real data results. Reported are SNPs selected in the seven prior-informative genes and their estimated odds ratios

Risk Gene	#SNPs	Prior Group Bridge		Group Bridge	
		SNP	Odds Ratio	SNP	Odds Ratio
DFNB31	4	rs10982246	0.950	–	–
		rs10982256	0.884	–	–
DGKH	6	rs619508	1.077	–	–
		rs9594703	1.143	–	–
EGFR	1	rs2293347	1.127	–	–
MYO5B	2	rs8098113	1.058	–	–
		rs17660456	1.119	–	–
NALCN	7	rs646773	0.891	–	–
		rs655365	1.274	–	–
		rs4771391	1.127	–	–
NXN	1	rs9892880	0.890	–	–
SORCS2	4	rs6446588	1.205	–	–
		rs16840358	1.237	–	–

Proof of Theorem 5

By the definition of a step, $\mathcal{A}(\lambda) \equiv \mathcal{A}(\lambda_0) = \mathcal{A}$ and the estimators $\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda)$ do not change their signs within a step. Then, the condition (7) in Theorem 1 implies that:

$$\text{(A.1)}\quad \mathbb{X}_{\mathcal{A}}^\top\{\mathbf{Y} - b'[\hat\beta_0(\lambda) + \mathbb{X}_{\mathcal{A}}^\top\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda)]\} = 2\lambda\|\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda)\|_1\mathbf{S}_{\mathcal{A}}.$$

Subtracting (A.1) evaluated at λ and λ_0 leads to:

$$\begin{aligned}&\mathbb{X}_{\mathcal{A}}^\top\{b'[\hat\beta_0(\lambda) + \mathbb{X}_{\mathcal{A}}^\top\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda)] - b'[\hat\beta_0(\lambda_0) + \mathbb{X}_{\mathcal{A}}^\top\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda_0)]\}\\ &= -2\mathbf{S}_{\mathcal{A}}[\lambda\|\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda)\|_1 - \lambda_0\|\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda_0)\|_1]\\ &= -2\mathbf{S}_{\mathcal{A}}\mathbf{S}_{\mathcal{A}}^\top[\lambda\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda) - \lambda_0\hat{\boldsymbol\beta}_{\mathcal{A}}(\lambda_0)].\end{aligned}$$

Proof of Theorem 7

Before presenting the main proof of Theorem 7, let us present a lemma that will be used in the proof.

Lemma 8. *Suppose* $n^{-1/2}\lambda_1 \to \lambda_1^* < \infty$ *and* $n^{-1/2}\lambda_2 \to \lambda_2^* < \infty$. *Assume that* $\boldsymbol\Sigma(\beta_{0,0}, \boldsymbol\beta_0)/n \to \boldsymbol\Sigma^*$ *where* $\boldsymbol\Sigma^*$ *is a positive definite matrix and* $\sup\{\|\boldsymbol\Sigma(b_0, \mathbf{b}) - \boldsymbol\Sigma(\beta_{0,0}, \boldsymbol\beta_0)\|_2 : \sqrt{n}\|\{b_0 - \beta_{0,0}, (\mathbf{b} - \boldsymbol\beta_0)^\top\}\|_2 \le \delta\} \to 0$ *for any* $\delta > 0$. *Then,* $\|\{\hat\beta_0 - \beta_{0,0}, (\hat{\boldsymbol\beta} - \boldsymbol\beta_0)^\top\}\|_2^2 = O_P(1/n)$.

Proof. By the definition of $(\hat\beta_0, \hat{\boldsymbol\beta})$, we have that

$$\text{(A.2)}\qquad \begin{aligned}&L(\hat\beta_0, \hat{\boldsymbol\beta}) - L(\beta_{0,0}, \boldsymbol\beta_0)\\ &\le \lambda_1\sum_{j\in\mathcal{J}_1} c_j(\|\boldsymbol\beta_{A_j,0}\|_1^2 - \|\hat{\boldsymbol\beta}_{A_j}\|_1^2)\\ &\quad + \lambda_2\sum_{j\in\mathcal{J}_2} c_j(\|\boldsymbol\beta_{A_j,0}\|_1^\gamma - \|\hat{\boldsymbol\beta}_{A_j}\|_1^\gamma).\end{aligned}$$

On the one hand, by Taylor expansion of the left-hand side of (A.2), $L(\hat\beta_0, \hat{\boldsymbol\beta}) - L(\beta_{0,0}, \boldsymbol\beta_0) = I_1 + I_2$ where

$$\begin{aligned}I_1 &= -\frac{1}{\sqrt{n}}\sum_{i=1}^n\{Y_i - b'(\beta_{0,0} + \mathbf{X}_i^\top\boldsymbol\beta_0)\}(1, \mathbf{X}_i^\top)\times\\ &\qquad \sqrt{n}\{\hat\beta_0 - \beta_{0,0}, (\hat{\boldsymbol\beta} - \boldsymbol\beta_0)^\top\}^\top,\\ I_2 &= \frac{1}{2n}\sum_{i=1}^n b''(\beta_{0,0} + \mathbf{X}_i^\top\boldsymbol\beta_0)[(1, \mathbf{X}_i^\top)\times\\ &\qquad \sqrt{n}\{\hat\beta_0 - \beta_{0,0}, (\hat{\boldsymbol\beta} - \boldsymbol\beta_0)^\top\}^\top]^2\{1 + o_P(1)\}.\end{aligned}$$

Similar to the proof of Theorem 3 in Jiang, He and Zhang [11], we can show that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n\{Y_i - b'(\beta_{0,0} + \mathbf{X}_i^\top\boldsymbol\beta_0)\}(1, \mathbf{X}_i^\top)\\ &\xrightarrow{d} N(\mathbf{0}, \boldsymbol\Sigma^*).\end{aligned}$$

This immediately leads to, for a large enough n,

$$\text{(A.3)}\quad \begin{aligned}|I_1| &= \|\sqrt{n}\boldsymbol\Sigma^{*1/2}\{\hat\beta_0 - \beta_{0,0}, (\hat{\boldsymbol\beta} - \boldsymbol\beta_0)^\top\}^\top\|_2 O_P(1)\\ &\le \frac{n}{4}\|\boldsymbol\Sigma^{*1/2}\{\hat\beta_0 - \beta_{0,0}, (\hat{\boldsymbol\beta} - \boldsymbol\beta_0)^\top\}^\top\|_2^2 + O_P(1).\end{aligned}$$

Further, when n is large enough,

$$\text{(A.4)}\quad I_2 = \frac{n}{2}\|\boldsymbol\Sigma^{*1/2}\{\hat\beta_0 - \beta_{0,0}, (\hat{\boldsymbol\beta} - \boldsymbol\beta_0)^\top\}^\top\|_2^2\{1 + o_P(1)\}.$$

On the other hand, for the right-hand side of (A.2), using the inequality that $b^r - a^r \le 2(b-a)b^{r-1}$ for $0 \le a \le b$ with $0 < r < 1$ or $r = 2$, we have that

$$\begin{aligned}&\|\boldsymbol\beta_{A_j,0}\|_1^r - \|\hat{\boldsymbol\beta}_{A_j}\|_1^r\\ &\le 2\|\boldsymbol\beta_{A_j,0}\|_1^{r-1}\|\hat{\boldsymbol\beta}_{A_j} - \boldsymbol\beta_{A_j,0}\|_1\\ &\le 2|A_j|^{1/2}\|\boldsymbol\beta_{A_j,0}\|_1^{r-1}\|\hat{\boldsymbol\beta}_{A_j} - \boldsymbol\beta_{A_j,0}\|_2.\end{aligned}$$

This leads to

$$\text{(A.5)}\quad \begin{aligned}&\lambda_1\sum_{j\in\mathcal{J}_1} c_j(\|\boldsymbol\beta_{A_j,0}\|_1^2 - \|\hat{\boldsymbol\beta}_{A_j}\|_1^2)+\\ &\lambda_2\sum_{j\in\mathcal{J}_2} c_j(\|\boldsymbol\beta_{A_j,0}\|_1^\gamma - \|\hat{\boldsymbol\beta}_{A_j}\|_1^\gamma)\\ &\le \lambda_1\sum_{j\in\mathcal{J}_1} c_j(\|\boldsymbol\beta_{A_j,0}\|_1^2 - \|\hat{\boldsymbol\beta}_{A_j}\|_1^2)+\\ &\quad \lambda_2\sum_{j\in\mathcal{J}_2:A_j\subseteq B_1} c_j(\|\boldsymbol\beta_{A_j,0}\|_1^\gamma - \|\hat{\boldsymbol\beta}_{A_j}\|_1^\gamma)\\ &\le 2\lambda_1\sum_{j\in\mathcal{J}_1} c_j|A_j|^{1/2}\|\boldsymbol\beta_{A_j,0}\|_1\|\hat{\boldsymbol\beta}_{A_j} - \boldsymbol\beta_{A_j,0}\|_2+\\ &\quad 2\lambda_2\sum_{j\in\mathcal{J}_2:A_j\subseteq B_1} c_j|A_j|^{1/2}\|\boldsymbol\beta_{A_j,0}\|_1^{\gamma-1}\|\hat{\boldsymbol\beta}_{A_j} - \boldsymbol\beta_{A_j,0}\|_2\\ &\le \lambda_1\eta_1\sqrt{\sum_{j\in\mathcal{J}_1}\|\hat{\boldsymbol\beta}_{A_j} - \boldsymbol\beta_{A_j,0}\|_2^2}+\\ &\quad \lambda_2\eta_2\sqrt{\sum_{j\in\mathcal{J}_2:A_j\subseteq B_1}\|\hat{\boldsymbol\beta}_{A_j} - \boldsymbol\beta_{A_j,0}\|_2^2}\\ &\le \max(\lambda_1\eta_1, \lambda_2\eta_2)\|\hat{\boldsymbol\beta} - \boldsymbol\beta_0\|_2,\end{aligned}$$

where $\eta_1 = 2\sqrt{\sum_{j\in\mathcal{J}_1} c_j^2 |A_j| \|\boldsymbol{\beta}_{A_j,0}\|_1^2}$ and $\eta_2 = 2\sqrt{\sum_{j\in\mathcal{J}_2:A_j\subseteq B_1} c_j^2 |A_j| \|\boldsymbol{\beta}_{A_j,0}\|_1^{2\gamma-2}}$ are bounded constants.

Combining (A.3)–(A.5) leads to

$$\begin{aligned}&\frac{n}{4}\|\boldsymbol{\Sigma}^{*1/2}\{\hat\beta_0-\beta_{0,0},(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^\top\}^\top\|_2^2\{1+o_P(1)\}\\ &\le \max(\lambda_1\eta_1,\lambda_2\eta_2)\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2+O_P(1).\end{aligned}$$

Therefore,

$$\begin{aligned}&\|\{\hat\beta_0-\beta_{0,0},(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^\top\}\|_2^2\\ &\le O\{\max(4\lambda_1\eta_1,4\lambda_2\eta_2)\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2/n\}+O_P(1/n)\\ &\le O\{\max(4\lambda_1\eta_1,4\lambda_2\eta_2)^2/(2n^2)\}+\frac{1}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2+O_P(1/n).\end{aligned}$$

Since $n^{-1/2}\lambda_1 \to \lambda_1^* < \infty$ and $n^{-1/2}\lambda_2 \to \lambda_2^* < \infty$, the above inequality implies that

$$\|\{\hat\beta_0-\beta_{0,0},(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^\top\}\|_2^2 = O_P(1/n).$$

So the proof is completed. □

Now we turn to prove part (a) of Theorem 7. Let's define an estimator $(\tilde\beta_0,\tilde{\boldsymbol{\beta}})$ so that $\tilde\beta_0=\hat\beta_0$, $\tilde{\boldsymbol{\beta}}_{B_1}=\hat{\boldsymbol{\beta}}_{B_1}$, and $\tilde{\boldsymbol{\beta}}_{B_2}=\mathbf{0}$. By the definitions of $(\hat\beta_0,\hat{\boldsymbol{\beta}})$ and $(\tilde\beta_0,\tilde{\boldsymbol{\beta}})$,

$$\text{(A.6)}\qquad \begin{aligned}&L(\tilde\beta_0,\tilde{\boldsymbol{\beta}})-L(\hat\beta_0,\hat{\boldsymbol{\beta}})\\ &\ge \lambda_1\sum_{j\in\mathcal{J}_1}c_j(\|\hat{\boldsymbol{\beta}}_{A_j}\|_1^2-\|\tilde{\boldsymbol{\beta}}_{A_j}\|_1^2)+\\ &\quad\lambda_2\sum_{j\in\mathcal{J}_2}c_j(\|\hat{\boldsymbol{\beta}}_{A_j}\|_1^\gamma-\|\tilde{\boldsymbol{\beta}}_{A_j}\|_1^\gamma)\\ &=\lambda_2\sum_{j:A_j\subseteq B_2}c_j\|\hat{\boldsymbol{\beta}}_{A_j}\|_1^\gamma,\end{aligned}$$

where the second equality follows from the assumption that $A_j\subseteq B_1$ for $j\in\mathcal{J}_1$.

Applying Taylor's expansion to the left-hand side of (A.6) leads to $L(\tilde\beta_0,\tilde{\boldsymbol{\beta}})-L(\hat\beta_0,\hat{\boldsymbol{\beta}})=I_1+I_2$, where

$$\begin{aligned}I_1 =&\ -\sum_{i=1}^n\{Y_i-b'(\hat\beta_0+\mathbf{X}_i^\top\hat{\boldsymbol{\beta}})\}(1,\mathbf{X}_i^\top)\times\\ &\ \{\tilde\beta_0-\hat\beta_0,(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^\top\}^\top,\\ I_2 =&\ \frac{1}{2}\sum_{i=1}^n b''(\beta_0^*+\mathbf{X}_i^\top\boldsymbol{\beta}^*)[(1,\mathbf{X}_i^\top)\times\\ &\ \{\tilde\beta_0-\hat\beta_0,(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})^\top\}^\top]^2\{1+o_P(1)\},\end{aligned}$$

with $(\beta_0^*,\boldsymbol{\beta}^*)$ lies between $(\hat\beta_0,\hat{\boldsymbol{\beta}})$ and $(\tilde\beta_0,\tilde{\boldsymbol{\beta}})$. For I_1, by KKT conditions,

$$\text{(A.7)}\quad \begin{aligned}|I_1| &= \left|\sum_{i=1}^n\{Y_i-b'(\hat\beta_0+\mathbf{X}_i^\top\hat{\boldsymbol{\beta}})\}\mathbf{X}_{i,B_2}^\top(\tilde{\boldsymbol{\beta}}_{B_2}-\hat{\boldsymbol{\beta}}_{B_2})\right|\\ &\le \gamma\lambda_2\sum_{j:A_j\subseteq B_2}c_j\|\hat{\boldsymbol{\beta}}_{A_j}\|_1^{\gamma-1}\sum_{k\in A_j}|\hat\beta_k|\\ &\le \gamma\lambda_2\sum_{j:A_j\subseteq B_2}c_j\|\hat{\boldsymbol{\beta}}_{A_j}\|_1^\gamma.\end{aligned}$$

For I_2, since $\boldsymbol{\Sigma}(\beta_{0,0},\boldsymbol{\beta}_0)\to\boldsymbol{\Sigma}^*$ and $\sup\{\|\boldsymbol{\Sigma}(b_0,\mathbf{b})-\boldsymbol{\Sigma}(\beta_{0,0},\boldsymbol{\beta}_0)\|_2:\sqrt{n}\|\{b_0-\beta_{0,0},(\mathbf{b}-\boldsymbol{\beta}_0)^\top\}\|_2\le\delta\}\to 0$ for any $\delta>0$, when n is large enough,

$$\text{(A.8)}\quad \begin{aligned}I_2 &= \frac{1}{2}\sum_{i=1}^n b''(\beta_0^*+\mathbf{X}_i^\top\boldsymbol{\beta}^*)(\mathbf{X}_{i,B_2}^\top\hat{\boldsymbol{\beta}}_{B_2})^2\{1+o_P(1)\}\\ &\ge \frac{n}{4}\lambda_{\min}(\boldsymbol{\Sigma}^*)\|\hat{\boldsymbol{\beta}}_{B_2}\|_2^2\{1+o_P(1)\}.\end{aligned}$$

Combining (A.6)–(A.8) leads to:

$$\text{(A.9)}\qquad \begin{aligned}&\frac{n}{4}\lambda_{\min}(\boldsymbol{\Sigma}^*)\|\hat{\boldsymbol{\beta}}_{B_2}\|_2^2\{1+o_P(1)\}\\ &\ge(1-\gamma)\lambda_2\sum_{j:A_j\subseteq B_2}c_j\|\hat{\boldsymbol{\beta}}_{A_j}\|_1^\gamma\\ &\ge(1-\gamma)\lambda_2\|\hat{\boldsymbol{\beta}}_{B_2}\|_1^\gamma\ge(1-\gamma)\lambda_2\|\hat{\boldsymbol{\beta}}_{B_2}\|_2^\gamma,\end{aligned}$$

where the second inequality follows from $a^\gamma+b^\gamma\ge(a+b)^\gamma$ for $a>0$, $b>0$, and $0<\gamma<1$. If $\hat{\boldsymbol{\beta}}_{B_2}\neq\mathbf{0}$, then (A.9) leads to:

$$\begin{aligned}(1-\gamma)\lambda_2 &\le \frac{n}{4}\lambda_{\min}(\boldsymbol{\Sigma}^*)\|\hat{\boldsymbol{\beta}}_{B_2}\|_2^{2-\gamma}\{1+o_P(1)\}\\ &= O_P\{n\times(1/\sqrt{n})^{2-\gamma}\}=O_P(n^{\gamma/2}).\end{aligned}$$

Therefore,

$$P(\hat{\boldsymbol{\beta}}_{B_2}\neq\mathbf{0})\le P\{(1-\gamma)\lambda_2\le O_P(n^{\gamma/2})\}\to 0,$$

because $\lambda_2/n^{\gamma/2}\to\infty$. The proof of part (a) of Theorem 7 is thus completed.

Finally, we prove part (b) of Theorem 7. Let $(\hat u_0,\hat{\mathbf{u}}) = \sqrt{n}\{\hat\beta_0-\beta_{0,0},(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}$. As $P(\hat{\boldsymbol{\beta}}_{B_2}=\mathbf{0})\to 1$, $\hat{\mathbf{u}}_{B_2}=\mathbf{0}$ with probability tending to one. Therefore, $(\hat u_0,\hat{\mathbf{u}}_{B_1})$ minimizes $V_n(u_0,\mathbf{u}_{B_1})$ where

$$\begin{aligned}&V_n(u_0,\mathbf{u}_{B_1})\\ &=L[\beta_{0,0}+\frac{1}{\sqrt{n}}u_0,(\boldsymbol{\beta}_{B_1,0}^\top+\frac{1}{\sqrt{n}}\mathbf{u}_{B_1}^\top,\mathbf{0}^\top)^\top]-\\ &\quad L(\beta_{0,0},\boldsymbol{\beta}_0)+\lambda_1\sum_{j\in\mathcal{J}_1}c_j\|\boldsymbol{\beta}_{A_j,0}+\frac{1}{\sqrt{n}}\mathbf{u}_{A_j}\|_1^2+\\ &\quad\lambda_2\sum_{j\in\mathcal{J}_2:A_j\subseteq B_1}c_j\|\boldsymbol{\beta}_{A_j,0}+\frac{1}{\sqrt{n}}\mathbf{u}_{A_j}\|_1^\gamma.\end{aligned}$$

Similar to the proof of Theorem 4 in Jiang, He and Zhang [11], it can be shown that

$$L[\beta_{0,0}+\frac{1}{\sqrt{n}}u_0,(\boldsymbol{\beta}_{B_1,0}^\top+\frac{1}{\sqrt{n}}\mathbf{u}_{B_1}^\top,\mathbf{0}^\top)^\top]-L(\beta_{0,0},\boldsymbol{\beta}_0)$$

$$\xrightarrow{d} -(u_0, \mathbf{u}_{B_1}^\top)\mathbf{Z} + \frac{1}{2}(u_0, \mathbf{u}_{B_1}^\top)\boldsymbol{\Sigma}_{1,1}^*(u_0, \mathbf{u}_{B_1}^\top)^\top,$$

where $\mathbf{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{1,1}^*)$. In addition, it is seen that

$$\lambda_1 \sum_{j \in \mathcal{J}_1} c_j (\|\boldsymbol{\beta}_{A_j,0} + \frac{1}{\sqrt{n}}\mathbf{u}_{A_j}\|_1^2 - \|\boldsymbol{\beta}_{A_j,0}\|_1^2) \to$$
$$2\lambda_1^* \sum_{j \in \mathcal{J}_1} c_j \|\boldsymbol{\beta}_{A_j,0}\|_1 \times$$
$$\sum_{k \in A_j \cap B_1} \{u_k \operatorname{sign}(\beta_{k,0}) I(\beta_{k,0} \neq 0) + |u_k| I(\beta_{k,0} = 0)\},$$
$$\lambda_2 \sum_{j \in \mathcal{J}_2 : A_j \subseteq B_1} c_j (\|\boldsymbol{\beta}_{A_j,0} + \frac{1}{\sqrt{n}}\mathbf{u}_{A_j}\|_1^\gamma - \|\boldsymbol{\beta}_{A_j,0}\|_1^\gamma) \to$$
$$\gamma\lambda_2^* \sum_{j \in \mathcal{J}_2} c_j \|\boldsymbol{\beta}_{A_j,0}\|_1^{\gamma-1} \times$$
$$\sum_{k \in A_j \cap B_1} \{u_k \operatorname{sign}(\beta_{k,0}) I(\beta_{k,0} \neq 0) + |u_k| I(\beta_{k,0} = 0)\}.$$

Therefore, $V_n(u_0, \mathbf{u}_{B_1}) \to V(u_0, \mathbf{u}_{B_1})$ in distribution. The proof is then completed by applying the continuous mapping theorem in Kim and Pollard [12].

Received 5 August 2019

REFERENCES

[1] Baum, A., Akula, N., Cabanero, M., Cardona, I., Corona, W., Klemens, B., Schulze, T., Cichon, S., Rietschel, M., Nöthen, M. et al. (2008a). A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Molecular Psychiatry* **13** 197–207.

[2] Baum, A., Hamshere, M., Green, E., Cichon, S., Rietschel, M., Noethen, M., Craddock, N. and McMahon, F. (2008b). Meta-analysis of two genome-wide association studies of bipolar disorder reveals important points of agreement. *Molecular Psychiatry* **13** 466.

[3] Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics* **71** 731–740. MR3402609

[4] Brzyski, D., Gossmann, A., Su, W. and Bogdan, M. (2019). Group slope–adaptive selection of groups of predictors. *Journal of the American Statistical Association* **114** 419–433. MR3941265

[5] Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

[6] Craddock, N. and Forty, L. (2006). Genetics of affective (mood) disorders. *European Journal of Human Genetics* **14** 660–668.

[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004). Least angle regression. *The Annals of Statistics* **32** 407–499. MR2060166

[8] Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L., Fan, J., Kirov, G., Perlis, R. H., Green, E. K. et al. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics* **40** 1056–1058.

[9] Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* **27**. MR3025130

[10] Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147

[11] Jiang, Y., He, Y. and Zhang, H. (2016). Variable selection with prior information for generalized linear models via the prior lasso method. *Journal of the American Statistical Association* **111** 355–376. MR3494665

[12] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics* 191–219. MR1041391

[13] Kwon, S., Ahn, J., Jang, W., Lee, S. and Kim, Y. (2017). A doubly sparse approach for group variable selection. *Annals of the Institute of Statistical Mathematics* **69** 997–1025. MR3689159

[14] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9** 356–369.

[15] Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 53–71. MR2412631

[16] Merikangas, K. R., Akiskal, H. S., Angst, J., Greenberg, P. E., Hirschfeld, R. M., Petukhova, M. and Kessler, R. C. (2007). Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of General Psychiatry* **64** 543–552.

[17] Ollila, H., Soronen, P., Silander, K., Palo, O., Kieseppä, T., Kaunisto, M., Lönnqvist, J., Peltonen, L., Partonen, T. and Paunio, T. (2009). Findings from bipolar disorder genome-wide association studies replicate in a Finnish bipolar family-cohort. *Molecular Psychiatry* **14** 351.

[18] Saxena, R., Elbers, C. C., Guo, Y., Peter, I., Gaunt, T. R., Mega, J. L., Lanktree, M. B., Tare, A., Castillo, B. A., Li, Y. R. et al. (2012). Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. *The American Journal of Human Genetics* **90** 410–425.

[19] Scott, L. J., Muglia, P., Kong, X. Q., Guan, W., Flickinger, M., Upmanyu, R., Tozzi, F., Li, J. Z., Burmeister, M., Absher, D. et al. (2009). Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proceedings of the National Academy of Sciences* **106** 7501–7506.

[20] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22** 231–245. MR3173712

[21] Sklar, P., Smoller, J., Fan, J., Ferreira, M., Perlis, R., Chambert, K., Nimgaonkar, V., McQueen, M., Faraone, S., Kirby, A. et al. (2008). Whole-genome association study of bipolar disorder. *Molecular Psychiatry* **13** 558–569.

[22] Smith, E. N., Bloss, C. S., Badner, J. A., Barrett, T., Belmonte, P. L., Berrettini, W., Byerley, W., Coryell, W., Craig, D., Edenberg, H. J. et al. (2009). Genome-wide association study of bipolar disorder in European American and African American individuals. *Molecular Psychiatry* **14** 755–763.

[23] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288. MR1379242

[24] Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis* **52** 5277–5286. MR2526593

[25] Wang, M. and Tian, G.-L. (2019). Adaptive group lasso for high-dimensional generalized linear models. *Statistical Papers* **60** 1469–1486. MR4017019

[26] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42** D1001–D1006.

[27] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical*

Society: Series B (Statistical Methodology) **68** 49–67. MR2212574

[28] Zhang, C. and Xiang, Y. (2016). On the oracle property of adaptive group Lasso in high-dimensional linear models. *Statistical Papers* **57** 249–265. MR3461958

[29] Zhao, P., Rocha, G., Yu, B. et al. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37** 3468–3497. MR2549566

Kai Li
Karyopharm Therapeutics Inc.
Newton
MA 02459
USA
E-mail address: lkustc24@gmail.com

Meng Mei
Department of Statistics
Oregon State University
Corvallis
OR 97331
USA
E-mail address: meim@oregonstate.edu

Yuan Jiang
Department of Statistics
Oregon State University
Corvallis
OR 97331
USA
E-mail address: yuan.jiang@stat.oregonstate.edu