Electronic Journal of Statistics
Vol. 12 (2018) 1988–2017
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1434

Selection by partitioning the solution paths

Yang Liu

Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109
e-mail: yliu23@fhcrc.org

and

Peng Wang

Department of Operations, Business Analytics and Information Systems, University of Cincinnati, Cincinnati, OH, 45220
e-mail: wangp9@ucmail.uc.edu

Abstract: The performance of penalized likelihood approaches depends profoundly on the selection of the tuning parameter; however, there is no commonly agreed-upon criterion for choosing the tuning parameter. Moreover, penalized likelihood estimation based on a single value of the tuning parameter suffers from several drawbacks. This article introduces a novel approach for feature selection based on the entire solution paths rather than the choice of a single tuning parameter, which significantly improves the accuracy of the selection. Moreover, the approach allows for feature selection using ridge or other strictly convex penalties. The key idea is to classify variables as relevant or irrelevant at each tuning parameter and then to select all of the variables which have been classified as relevant at least once. We establish the theoretical properties of the method, which requires significantly weaker conditions than existing methods in the literature. We also illustrate the advantages of the proposed approach with simulation studies and a data example.

MSC 2010 subject classifications: Primary 62F07; secondary 62J07, 62J86.
Keywords and phrases: Penalized likelihood, lasso, variable/feature selection, solution paths, AIC/BIC, cross-validation, tuning.

Received January 2018.

Contents

1 Introduction 1989
2 Selection by partitioning the solution paths 1991
3 Notations and algorithm 1993
  3.1 Notations 1993
  3.2 Algorithm of SPSP 1995
4 Consistent feature selection 1997
5 Simulation studies 2002
6 Data analysis 2004
7 Discussion 2008

8 Code 2010
A Appendix 2010
A.1 Technical proofs 2010
A.2 Simulation studies on the sensitivity of SPSP to the value of R 2012
A.3 Simulation study on the Gaussian graphical model 2014
References 2014

1. Introduction

The penalized likelihood approach has been very popular for feature selection problems. Under the currently used framework, one first needs to compute the solution paths and then choose a tuning parameter based on a certain criterion. The solution yielded with the chosen tuning parameter is considered to provide the final estimates of the parameters. The problem of choosing a proper tuning parameter is notoriously difficult. In this paper, we propose to select features by utilizing the entire solution paths rather than searching for a single value of the tuning parameter.

Consider the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon};\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is an n-dimensional response vector, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ is a p-dimensional vector of regression coefficients, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p})$ is an $n \times p$ design matrix, and $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is an n-dimensional vector of independent and identically distributed random errors. We assume that all of the variables in $\mathbf{X}$ are standardized, so that the coefficients in $\boldsymbol{\beta}^*$ are on the same scale. For the linear regression problems described in (1.1), the penalized likelihood approach is equivalent to the penalized least squares regression, and the regression coefficients are estimated by minimizing the following objective function:

$$\frac{1}{n}\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} J(|\beta_j|), \tag{1.2}$$

where $J(\cdot)$ is a penalty function which controls the number of non-zero coefficients and $\lambda > 0$ is a tuning parameter. For the penalty function $J(\cdot)$, one could use a convex penalty like the lasso (Tibshirani, 1996) or the adaptive lasso (Zou, 2006) or one of the non-convex penalties such as the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the minimax concave penalty (MCP) (Zhang, Li and Tsai, 2010), or the truncated l_1 penalty (Shen, Pan and Zhu, 2012), among others.

Some general selection criteria for the tuning parameter λ include cross-validation (CV), generalized cross-validation (GCV), the Akaike information criterion (AIC) (Akaike, 1973), the Bayesian information criterion (BIC) (Schwarz, 1978), and the generalized information criterion (Atkinson, 1980). Chen and Chen (2008) pointed out that these criteria usually identify too many irrelevant

Table 1
The general selection criteria in the statistics and machine learning literature in major journals.

Criterion	Method	References
CV	Adaptive lasso	Zou (2006), JASA
	Fused lasso	Tibshirani et al. (2005), JRSSB
	Truncated l_1 penalty	Shen, Pan and Zhu (2012), JASA
GCV	Lasso	Tibshirani (1996), JRSSB
	SCAD	Fan and Li (2001), JASA
BIC	Lasso	Wang, Li and Tsai (2007), JRSSB
	Lasso	Yuan and Lin (2007), Biometrika
EBIC	Tilting	Cho and Fryzlewicz (2012), JRSSB
	Group lasso	Huang, Horowitz and Wei (2010), AOS

features when the number of variables is large. This phenomenon has also been described by Broman and Speed (2002), Siegmund (2004), and Bogdan, Ghosh and Doerge (2004) in their studies of quantitative loci mapping. Chen and Chen (2008) proposed the extended BIC (EBIC), which promotes model sparsity by adjusting BIC with an additional penalty term for the growing number of parameters in the model. Recently, Sun, Wang and Fang (2013) developed a new technique via variable selection stability, which directly focuses on the selection of informative variables. Fan and Tang (2013) also proposed to select the tuning parameter by optimizing the generalized information criterion with an appropriate penalty depending on the dimensions in the model.

Although the above criteria have been well studied for over a decade, there is no concurrence of opinion regarding which criterion to employ for choosing the tuning parameter. As examples, see Table 1 for a short list of publications in major statistics and machine learning journals and the different criteria they use. In fact, the feature selection procedure currently in use, which utilizes only one chosen value for the tuning parameter, suffers from the unavoidable drawback that it is often impossible to correctly identify all the features, no matter which criterion is used.

To overcome this drawback, we develop an innovative, intuitive approach which avoids choosing a tuning parameter. Instead, our proposed approach utilizes information from the entire solution paths to improve the selection accuracy. Thus we named the proposed procedure *selection by partitioning the solution paths* (SPSP). Besides improvement in selection accuracy, the SPSP approach can also achieve selection consistency under a weaker condition compared to the currently used framework. This is because SPSP does not require selection consistency at any specific values of the tuning parameter. As a matter of fact, we do not even require the estimation of coefficients shrunk to zeros. This brings another major advantage of the SPSP approach: we can achieve feature selection with an l_2 penalty or ridge regression. It is well known that ridge regression is more stable and handles collinearity better compared to the lasso. Now with the SPSP approach, we can enjoy these nice properties of the ridge regression without sacrificing capabilities of feature selection. Moreover, the minimizer of l_2 penalized likelihood function often has an explicit solution.

Therefore SPSP also greatly reduces the efforts in algorithm development in presence of complicated likelihood functions.

The SPSP procedure is related to the stability selection approach proposed in the seminal discussion paper by Meinshausen and Bühlmann (2010). Their approach is based on the probabilities that variables will be selected, and the probabilities are obtained from a generic sub-sampling approach. Therefore, the stability selection approach does not require the selection of a tuning parameter either. However, unlike the proposed SPSP procedure, stability selection does not work with ridge regression. Moreover, the computational cost of the SPSP procedure is only a tiny fraction of that for stability selection, since no sub-sampling is involved. Finally, we find from simulation studies that stability selection tends to select too few variables and therefore produces a significantly higher false negative rate than SPSP. This work is also remotely related to Bayesian variable selection approaches, where the tuning parameters or candidate models are assigned a prior distributions and the posterior distributions of the models are evaluated. See, for example, Hoeting et al. (1999); Raftery, Madigan and Hoeting (1997); Posada and Buckley (2004); Barbieri and Berger (2004).

The rest of this article is organized as follows. Section 2 provides a simulated example to motivate the problem. Section 3 introduces the SPSP approach. Section 4 discusses the selection consistency of the SPSP procedure. Section 5 presents the results from various simulation examples. Section 6 provides an application of SPSP in a cancer study to detect the significant genes for glioblastomas, and Section 7 discusses the advantages and future potential of this work. Appendix A provides technique proofs and additional simulation studies.

2. Selection by partitioning the solution paths

In this section, we will illustrate the advantage and the idea of the proposed SPSP procedure with an example from model (1.1) in Wang et al. (2011), where $p = 40$, $\beta_1^* = \cdots = \beta_5^* = 3$, and $\beta_6^* = \cdots = \beta_{10}^* = -2$. The rest of the coefficients $\beta_j^*, j = 11, \ldots, 40$, are all zeros. The entries for the variables $\mathbf{x}_j, j = 1, \ldots, p$, are generated from a multivariate Gaussian distribution whose marginals are the standard normal distribution. The pairwise correlation between the first 10 variables $\mathbf{x}_j, j = 1, \ldots, 10$, is 0.9. The remaining 30 variables $\mathbf{x}_j, j = 11, \ldots, 40$, are mutually independent and they are also independent of the first 10 variables. Furthermore, we generate errors from the normal distribution with mean 0 and standard deviation 3. The sample size is set to be $n = 50$.

We compute the solution paths and plot them in Figure 1. The dashed lines represent the solution paths for the non-zero coefficients and the solid lines represent those for the zero ones. The tuning parameters chosen by the twofold cross-validation, generalized cross-validation , AIC, BIC, and the extended BIC are shown by the vertical lines. Here, "BEST" is the tuning parameter which minimizes the number of incorrect selections (false positives + false negatives). We observe that cross-validation, generalized cross-validation, AIC, and BIC

tend to select too many spurious variables and that the extended BIC tends to drop most of the important variables. Even for the "BEST" selection, the model excludes many non-zero coefficients. The problem is more evident when we focus on the three lines in the right panel of Figure 1. Here the two dotted lines (1 and 3) are the solution paths for two non-zero coefficients, while the solid line (2) is the solution path for a zero coefficient. Apparently, selecting a small λ, as AIC, BIC, and generalized cross-validation do, misleads us into identifying all three coefficients as non-zero. On the other hand, a large λ, as selected in cross-validation, the extended BIC, and "BEST", incorrectly shrinks both coefficients of the important variables to zeros. In fact, it is impossible to correctly identify all three features no matter which value of the tuning parameter we choose, although one can likely see the differences between the three features by visual inspection. For this example, we also provide the solution paths using elastic-net penalty in the left panel of Figure 3, which shows that these traditional criteria for selecting the tuning parameter suffer from the same problem.

The current restriction of utilizing just one tuning parameter can seriously reduce the accuracy of the feature selection in general, since solution paths like those in Figure 1 happen quite frequently, and for any penalty we may employ. This is especially true when there are large correlations among the variables or the dimensions of the features are extremely high.

Such deficiencies as described above motivate us to develop an approach that allows for collectively using evidence from the entire solution paths, as opposed to letting results from a single value of λ dictate everything. We achieve this by first dividing estimates from each value of λ into two clusters. Then we combine the cutoff points of all such clusters to form a boundary curve to partition the solution paths into two regions, as shown by the red curves in Figure 2. We call the region inside the two red curves the zero region and the region outside the two curves the non-zero region. Finally, we choose all variables which have been identified as relevant variables for at least one value of λ as the important features. We consider a feature to be unimportant if its solution path never goes out of the zero region.

It can readily be observed in Figure 2 that the SPSP procedure correctly selects 9 of the 10 relevant variables and drops all irrelevant ones, outperforming the results for any single value of λ in terms of selection accuracy. As a result, our approach can correctly identify the relevance of features like the three in Figure 1, labeled 1, 2, and 3. Additionally, we apply this procedure on the solution paths of elastic-net for this example. As seen in the right panel of Figure 3, the procedure can improve the selection accuracy for the elastic-net as well. Another advantage of SPSP is that it does not require the coefficients of the unimportant variables to be shrunk to zeros; therefore, it allows us to carry out feature selection with only a ridge regression.

The proposed SPSP procedure considers a feature important even if its solution path enters the non-zero region only once. The strategy may seem aggressive in identifying relevant variables. This is because we start the SPSP process rather conservatively, in the sense that we consider every variable "unimportant" for the smallest value of λ, when we initiate the partitioning process. The

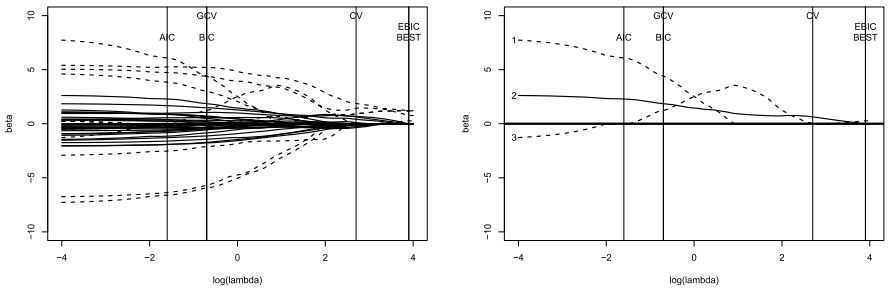

FIG 1. *Left: The lasso solution paths for the simulated example. The dashed lines are the paths of the* 10 *non-zero coefficients, while the black lines are the paths of the* 30 *zero coefficients The vertical lines represent the tuning parameters selected by different criteria. Right: The lasso solution paths for the non-zero coefficients, 1 and 3, and the zero coefficient, 2. Here CV is cross-validation, GCV is generalized cross-validation and EBIC is extended BIC.*

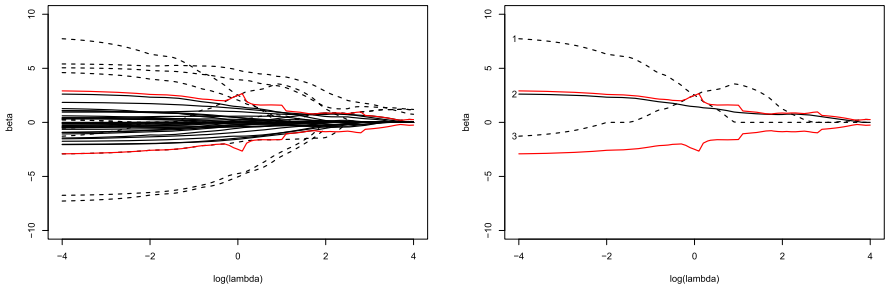

FIG 2. *Left: Partitions of the lasso solution paths of the same simulated example. Right: Partitions of the lasso solution paths for the non-zero coefficients, 1 and 3, and the zero coefficient, 2.*

clustering at larger values of λ depends on the results for the previous λ. Thus, the SPSP procedure combines a conservative starting point with an aggressive selection strategy to optimize the selection accuracy.

3. Notations and algorithm

3.1. Notations

Consider the penalized least squares problem in (1.2). We denote the index set for the non-zero coefficients as $S = \{j : \beta_j^* \neq 0\}$, with $s = |S|$, and the index set for the zero coefficients as $S^c = \{j : \beta_j^* = 0\}$. The goal of feature selection is to correctly recover this sparsity pattern from the noisy observations in the model and estimate S. Once we specify the penalty function $J(\cdot)$ in (1.2), we can compute the solution paths on a grid for the tuning parameters $\lambda_1 < \lambda_2 < \cdots < \lambda_k$. In practice, the grid is usually equi-distant on the log scale (Bühlmann and van de Geer, 2011; Shen, Pan and Zhu, 2012). For each

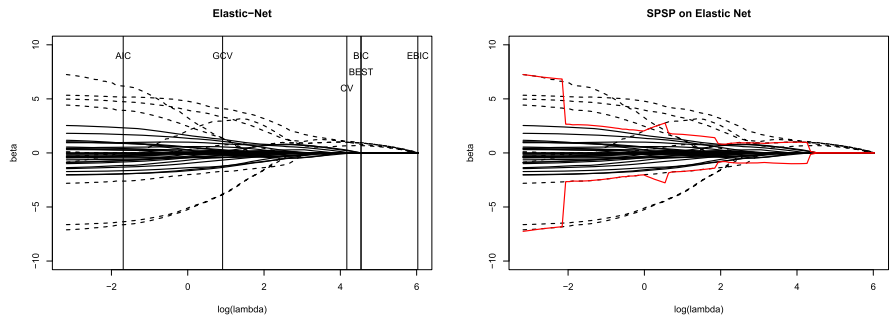

FIG 3. *Left: The elastic net solution paths for the simulated example. The dashed lines are the paths of the non-zero coefficients, while the black lines are the paths of the zero coefficients The vertical lines represent the tuning parameters selected by different criteria. Right: Partitions of the elastic net solution paths of the example using SPSP.*

λ_k, we obtain a vector $\hat{\boldsymbol{\beta}}^{(\boldsymbol{k})} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})^T$ of the penalized least squares estimators. A variable is more likely to be identified as relevant if its estimator is farther away from 0, regardless of the sign of the estimator. Therefore, we use the absolute values $\hat{\boldsymbol{b}}^{(\boldsymbol{k})} = (\hat{b}_1^{(k)}, \ldots, \hat{b}_p^{(k)})^T = (|\hat{\beta}_1^{(k)}|, \ldots, |\hat{\beta}_p^{(k)}|)^T$ of the estimators.

In general, variables with larger values of $\hat{b}_j^{(k)}$'s are more likely to be important. We are therefore interested in finding a proper cutoff point $T_k = T(\lambda_k)$ such that the estimated relevant set $\hat{S}_k$ and irrelevant set $\hat{S}_k^c$ at $\lambda = \lambda_k$ are derived as $\hat{S}_k = \{j : \hat{b}_j^{(k)} > T_k\}$ and $\hat{S}_k^c = \{j : \hat{b}_j^{(k)} \leq T_k\}$. To obtain $(T_k, \hat{S}_k, \hat{S}_k^c)$ for each λ_k, we sort the absolute values $\hat{b}_1^{(k)}, \ldots, \hat{b}_p^{(k)}$ in ascending order: $\hat{b}_{(1)}^{(k)} \leq \cdots \leq \hat{b}_{(p)}^{(k)}$, where $\hat{b}_{(j)}^{(k)}$ is the jth order statistics of $\hat{b}_1^{(k)}, \ldots, \hat{b}_p^{(k)}$. Then we define the adjacent distances between these ordered values as

$$D_j^{(k)} = \hat{b}_{(j)}^{(k)} - \hat{b}_{(j-1)}^{(k)}, j = 1, \ldots, p. \tag{3.1}$$

Note that $D_1^{(k)}$ is the adjacent distance between $\hat{b}_{(1)}^{(k)}$ and 0, as we define $\hat{b}_{(0)}^{(k)} = 0$ for convenience. Letting $\hat{s}_k = |\hat{S}_k|$ be the number of variables in the estimated relevant set $\hat{S}_k$, there are $p - \hat{s}_k$ variables in the estimated irrelevant set $\hat{S}_k^c$. Hereafter, we define the gap between $\hat{S}_k$ and $\hat{S}_k^c$ as the adjacent distance between $\hat{b}_{(p-\hat{s}_k)}^{(k)}$ and $\hat{b}_{(p-\hat{s}_k+1)}^{(k)}$, i.e. $D(\hat{S}_k, \hat{S}_k^c) = D_{p-\hat{s}_k+1}^{(k)} = \hat{b}_{(p-\hat{s}_k+1)}^{(k)} - \hat{b}_{(p-\hat{s}_k)}^{(k)}$. Further, let $D_{\max}(\hat{S}_k) = \max\{D_j^{(k)} : j > p - \hat{s}_k + 1\}$ be the largest adjacent distance in $\hat{S}_k$, let $D_{\max}(\hat{S}_k^c) = \max\{D_j^{(k)} : j < p - \hat{s}_k + 1\}$ be the largest adjacent distance in $\hat{S}_k^c$, and let

$$D_{\max 2}(\hat{S}_k^c) = \max\{D_j^{(k)} : j < \tilde{j}, D_{\tilde{j}}^{(k)} = D_{\max}(\hat{S}_k^c)\}$$

be the largest adjacent distance under $D_{\max}(\hat{S}_k^c)$.

3.2. Algorithm of SPSP

To identify an important feature $j, j \in S$ correctly without introducing any spurious variables, the estimate $\hat{b}_j^{(k)}$ has to be larger than the estimates of zero coefficients at some λ_k. But unlike the currently used methods of using just one tuning parameter, we do not require all of the estimates $\hat{b}_j^{(k)}$ "concurrently" larger than all the estimates of zero coefficients at a certain λ_k. Rather we just need $\hat{b}_j^{(k)}$ to be relatively larger at some λ_k for variable j to be selected. That is, even if $\hat{b}_j^{(k)}$ and $\hat{b}_{j'}^{(k)}$, $j, j' \in S$ take larger values (compared to the estimates of zero coefficients) at different λ_k parameters, we can still identify both j and j' without the cost of adding any false positive signals. The idea is to distinguish the larger estimates from the smaller ones at each λ_k and consider those variables with the larger estimates as $\hat{S}_k$.

The major issue therefore is to find a proper way to separate the larger and smaller estimates at each λ_k. It is inappropriate to use a constant threshold because for larger λ_k all the estimates might be small. However, the order of $\hat{b}_j^{(k)}, j \in S$ is still likely larger than that of estimates of zero coefficients. Following this intuition, we can set the threshold T_k at λ_k as $\{T_k : \hat{b}_j^{(k)} \leq C\hat{b}_l^{(k)}, \hat{b}_j^{(k)} > C\hat{b}_s^{(k)}$ for any $\hat{b}_j^{(k)}, \hat{b}_l^{(k)} > T_k, \hat{b}_s^{(k)} \leq T_k\}$, where C is a certain constant. This way, we consider an estimate $\hat{b}_j^{(k)}$ relatively larger if $\hat{b}_j^{(k)} > C\hat{b}_s^{(k)}$ for any smaller estimates $\hat{b}_s^{(k)}$ and guarantee that all the larger estimates are of the same order. As we can imagine, such a threshold happens where the adjacent distance (3.1) is relatively large. Therefore, instead of comparing all the $\hat{b}_j^{(k)}$ pairwise to find T_k, we can equivalently try to find an adjacent distance that is large enough to separate $\hat{S}_k$ and $\hat{S}_k^c$.

In principle, $D(\hat{S}_k, \hat{S}_k^c)$, the gap between $\hat{S}_k$ and $\hat{S}_k^c$, should be sufficiently large to separate the irrelevant features from the important ones if 1. all $\hat{\beta}_j^{(k)}$ above the gap $D(\hat{S}_k, \hat{S}_k^c)$ are of the same order, 2. all $\hat{\beta}_j^{(k)}$ above the gap $D(\hat{S}_k, \hat{S}_k^c)$ are of higher order than those below the gap. It is not hard to verify that these are equivalent to the following statements in terms of adjacent distances: 1. the adjacent distances above the gap $D(\hat{S}_k, \hat{S}_k^c)$ are not of higher order than the gap; 2. the adjacent distances below the gap $D(\hat{S}_k, \hat{S}_k^c)$ are of lower order than the gap. For that reason, we consider $D(\hat{S}_k, \hat{S}_k^c)$ to be large enough if it meets the following two criteria:

$$\frac{D_{\max}(\hat{S}_k)}{D(\hat{S}_k, \hat{S}_k^c)} \leq R, \tag{3.2}$$

$$\frac{D(\hat{S}_k, \hat{S}_k^c)}{D_{\max}(\hat{S}_k^c)} > R, \tag{3.3}$$

where R is a constant which we can estimate from the data.

Based on this principle, we develop the SPSP algorithm as follows:

Algorithm 1: Selection by Partitioning the Solution Paths (SPSP)

1 At λ_1, set the initial values to $T_1 = \infty, \hat{S}_1 = \emptyset$, and $\hat{S}_1^c = \{1, \ldots, p\}$, and estimate the constant R as $R = D_{\max}(\hat{S}_1^c)/D_{\max 2}(\hat{S}_1^c)$.

2 At each λ_k, estimate $T_k, \hat{S}_k$, and $\hat{S}_k^c$ from $T_{k-1}, \hat{S}_{k-1}, \hat{S}_{k-1}^c$, and $\hat{\boldsymbol{b}}^{(k)}$.

2.1 Update $T_k = \max_{j \in \hat{S}_{k-1}^c} \hat{b}_j^{(k)}, \hat{S}_k = \{j : \hat{b}_j^{(k)} > T_k\}, \hat{S}_k^c = \{j : \hat{b}_j^{(k)} \leq T_k\}$;

2.2 Calculate $D_1^{(k)}, \ldots, D_p^{(k)}$. Further, obtain $D_{\max}(\hat{S}_k^c), D_{\max 2}(\hat{S}_k^c)$, and $D(\hat{S}_k, \hat{S}_k^c)$.

2.3 If $D(\hat{S}_k, \hat{S}_k^c) \leq R \times D_{\max}(\hat{S}_k^c)$ and $D_{\max}(\hat{S}_k^c) > R \times D_{\max 2}(\hat{S}_k^c)$, $D_{\max}(\hat{S}_k^c)$ is the new gap between S_k and S_k^C, i.e. $D(\hat{S}_k, \hat{S}_k^c) = D_{\max}(\hat{S}_k^c)$. Therefore we also update

$$T_k = \hat{b}_{(\tilde{j}-1)}^{(k)}, \hat{S}_k = \{j : \hat{b}_j^{(k)} > T_k\}, \hat{S}_k^c = \{j : \hat{b}_j^{(k)} \leq T_k\},$$

where $\tilde{j}$ is the location of $D_{\max}(\hat{S}_k^c)$ i.e. $D_{\tilde{j}}^{(k)} = D_{\max}(\hat{S}_k^c)$.

Otherwise $T_k, \hat{S}_k, \hat{S}_k^c$ remain unchanged as in Step 2.1.

3 Make $k = k + 1$ and repeat Step 2 until $k = K$.

4 Identify the union of all $\hat{S}_k$ as the index set for our selected relevant variables, i.e. $\hat{S} = \bigcup_{k=1}^{K} \hat{S}_k$.

Remark 3.1. Principles (3.2) and (3.3) are implemented in Step 2.3. They are equivalent to the claim that the estimators of the non-zero coefficients should be of the same order and they should have a higher order than the estimators of the zero-coefficients. Instead of comparing every pair of the estimators, we just use adjacent distances for simplicity of computation. Therefore, finding the proper T_k now transforms to finding an adjacent distance which is large enough—i.e. which satisfies (3.2) and (3.3)—to be the gap between the estimators for the zero coefficients and those for the non-zero ones. The constant R, which we obtain from the data in Step 1, is used to decide whether $D_{\max}(\hat{S}_k^c)$ is large enough to replace the current $D(\hat{S}_k, \hat{S}_k^c)$. As a matter of fact, our final selection results are not so sensitive to the value of R, as evidenced by the simulation study conducted in Appendix A.2.

Remark 3.2. In Step 2 we find, for each λ_k, the cutoff point T_k, the location of the gap which distinguishes the relevant and irrelevant variables, based on the results from λ_{k-1}. This not only simplifies the computation process, but also makes the boundary line $T_k = T(\lambda_k)$ smoother to avoid unstable selection results. Specifically, in Step 2.1, we first use the largest estimated coefficients among those identified as "zero coefficients" for λ_{k-1} as the current boundary. This takes care of the case where some coefficients in $\hat{S}_{k-1}$ become small and enter into the zero region at λ_k. In Step 2.2 and Step 2.3, we decide whether

any adjacent distances within $\hat{S}_k^c$ are large enough to be considered as the new gap between the zero and non-zero coefficients. This handles the scenario where there are too few variables in $\hat{S}_k$, so that a "large" gap still exists.

Remark 3.3. When the algorithm starts, with λ_1, we use the initial values set in Step 1, with all variables considered "irrelevant". In other words, we are conservative about identifying the relevant variables at the start of the procedure. This is because we implement an aggressive selection strategy in Step 4 by using the union of all $\hat{S}_k$'s as our estimator $\hat{S}$ for the set of the relevant variables. This effectively balances out the very conservative initial setup to achieve the best selection accuracy. Another reason for the conservative initial setup is that the estimations at the small λ_k parameters are usually unstable, as there might be too many non-zeros in $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ since the shrinkage effects of the penalty are minimal. As a result, even the estimation of a zero coefficient can possibly be rather large.

Once we identify the index set of all the relevant variables $\hat{S}$, we estimate the regression parameters $\hat{\boldsymbol{\beta}}_{\hat{S}}$ in a model that only includes the features that have been selected, $\mathbf{y} = \mathbf{X}_{\hat{S}}\boldsymbol{\beta}_{\hat{S}} + \boldsymbol{\varepsilon}$, where $\mathbf{X}_{\hat{S}} = (\mathbf{x_j})_{j\in\hat{S}}$ and $\boldsymbol{\beta}_{\hat{S}} = (\beta_j)^T_{j\in\hat{S}}$. In most cases, the number of features in $\hat{S}$ is smaller than the sample size, so we just use the least squares estimator as $\hat{\boldsymbol{\beta}}_{\hat{S}}$. If the number of selected variables is larger than the sample size, we can use a ridge regression with a small shrinkage factor.

One unique advantage of this procedure is that it can be applied not only to penalties like lasso, adaptive lasso, SCAD, and MCP; it can also be applied to penalties which do not produce the sparse solutions, such as the l_2 penalty. Therefore, the procedure can greatly reduce computational complexity for feature selection problems, since strictly convex penalties like a ridge are easier to solve and their estimators have an explicit form.

In addition, the SPSP algorithm can be easily extended to handle selection problems in a wide range of models, including graphical models, generalized linear models, and Cox's proportional hazards models. The penalized likelihood approach, which obtains a sparse estimate by solving an objective function consisting of a likelihood and a penalty function, is usually applied in these models. Consequently, we can apply the SPSP algorithm to penalized likelihood estimators in a similar fashion. A simulation example using SPSP on Gaussian graphical models is provided in Appendix A.3.

4. Consistent feature selection

In this section, we discuss the theoretical properties of the SPSP procedure with lasso for feature selection in a modern high-dimensional regime where $p >> n$. Here we limit our efforts to linear regression, although the SPSP procedure is generally applicable to other selection problems as well. The technical proofs of all the lemmas and theorems in this section are available in Appendix A.1.

"Consistent variable selection" for a procedure refers to the following property of its estimator $\hat{S}$:

$$P(\hat{S} = S) \to 1 \text{ as } n \to \infty.$$

In most of the existing literature, it is only possible to achieve feature selection consistency if the tuning parameter is restricted to a specific interval. Moreover, the widths of such intervals are usually so small that they converge to 0—see, for example, Fan and Li (2001), Fan and Peng (2004), Zhao and Yu (2006), and Zou (2006). It is also commonly recognized that under the high-dimensional setting, $p >> n$, the Gram matrix $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ is degenerate, which raises many difficulties in controlling the values of the lasso estimator. Therefore, some conditions on the design matrix are always required to establish the consistency of the feature selection. The most typical condition is probably the following irrepresentable condition (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Zhao and Yu, 2006; Zou, 2006):

$$\left|\frac{1}{n}\mathbf{X}_{S^c}^T\mathbf{X}_S\left(\frac{1}{n}\mathbf{X}_S^T\mathbf{X}_S\right)^{-1}\right| \leq 1-\eta,$$

where η is a positive constant. Zhao and Yu (2006) showed that this condition is sufficient and almost necessary for lasso to be selection consistent. However, the condition is restrictive and difficult to verify in practice. Here, we will first show that the SPSP procedure is selection consistent under either a much weaker compatibility condition (Bühlmann and van de Geer, 2011) or the restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009) if we can bound the tuning parameter to an interval of constant width, rather than one that is converging to zero. We further show that under a weak identifiability condition, the SPSP procedure achieves selection consistency for almost all values of the tuning parameter, i.e. the entire solution paths. The weak identifiability condition is still weaker than the irrepresentable condition.

We first introduce the following compatibility condition:

Assumption 4.1. ***Compatibility Condition*** *(Bühlmann and van de Geer, 2011; van de Geer, 2007). For some constant $\phi > 0$ and for any vector $\boldsymbol{\zeta}$ satisfying $||\boldsymbol{\zeta}_{S^c}||_1 \leq 3||\boldsymbol{\zeta}_S||_1$, the following compatibility condition holds:*

$$||\boldsymbol{\zeta}_S||_1^2 \leq \left(\boldsymbol{\zeta}^{\boldsymbol{T}}\hat{\Sigma}\boldsymbol{\zeta}\right) s/\phi^2,$$

where $s = |S|$ is the dimension of $\boldsymbol{\beta}_S$.

The compatibility condition is based on the fact that the bias of the lasso estimator $\boldsymbol{\zeta} = \hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*$ satisfies $||\boldsymbol{\zeta}_{S^c}||_1 \leq 3||\boldsymbol{\zeta}_S||_1$ with a probability close to 1 (Bickel, Ritov and Tsybakov, 2009; Bühlmann and van de Geer, 2011). Hence we can restrict ourselves to such vectors in the condition. Several similar assumptions have also been proposed to establish the consistency property of the lasso, such as the restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009), the restricted isometry condition (Candes and Tao, 2005), and the coherence condition (Bunea, Tsybakov and Wegkamp, 2007). The relations among these conditions can be found in Bühlmann and van de Geer (2011). Under the compatibility condition, we can bound both the bias and the prediction error of the lasso:

Lemma 4.1. *(Bühlmann and van de Geer, 2011) Suppose that the compatibility condition holds, and let* $\lambda_0 = 2\sigma\sqrt{\frac{t^2+2\log p}{n}}$ *for any* $t > 0$*. Then for* $\lambda \geq 2\lambda_0$*, we have*

$$\frac{1}{n}||\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||_2^2 + \lambda||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leq \frac{4\lambda^2 s}{\phi^2}$$

with a probability of at least $1 - 2e^{-t^2/2}$*.*

This lemma implies the bound for the prediction error and and the following bound for the l_1-error of the lasso estimator:

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leq \frac{4\lambda s}{\phi^2}.$$

The compatibility condition required to bound the above errors is substantially weaker than the irrepresentable condition, which is necessary for achieving consistent variable selection under the currently used framework of choosing a single λ. Bühlmann and van de Geer (2011) showed that the irrepresentable condition actually implies the compatibility condition.

With the proposed SPSP procedure, we can accomplish consistent variable selection without the irrepresentable condition. This is because at each λ, we cluster the lasso estimators into two groups rather than labeling all the variables with non-zero coefficient estimates as important features. Consequently, we only need to bound the bias of the lasso estimators rather than shrink some coefficients to zeros. In what follows, we will first introduce some necessary notations and then present Theorem 4.1. The theorem shows that when λ is not too large, the SPSP procedure identifies the true relevant set S with a probability close to 1 with only the compatibility condition.

Let $\delta_\lambda = \frac{4\lambda s}{\phi^2}$ and $\delta_0 = \delta_{\lambda_0}$, where λ_0 is defined as in Lemma 4.1. We first sort the absolute values of the true non-zero coefficients $\{b_j^* = |\beta_j^*|, j \in S\}$ in ascending order to get $b_{(1)}^*, \ldots, b_{(s)}^*$, and we define the true adjacent distances as $D_0 = 0, D_1 = b_{(1)}^*, D_2 = b_{(2)}^* - b_{(1)}^*, \ldots, D_s = b_{(s)}^* - b_{(s-1)}^*$. Let $C_0 = \sqrt{\frac{D_{\max}+\delta_0}{\delta_0}} - 1$, $D_{\max} = \max_{1\leq i\leq p}\{D_i\}$, and

$$C = \frac{D_{\max}}{\min\{b_i^* : b_i^* > (2 + C_0)\delta_0\}}.$$

Moreover, let

$$C_{under}^i = \frac{D_i}{\max\{D_{i'} : i' < i\}}$$

for $i = 2, \ldots, s$ and $C_{under}^1 = \infty$, and

$$C_{upper}^i = \frac{D_i}{\max\{D_{i'} : i' > i\}}$$

for $i = 1, \ldots, s-1$ and $C_{upper}^s = \infty$. In addition, let $R = 1 + C$. We also sort the absolute values of the estimators from lasso in ascending order to get $\hat{b}_{(1)}, \ldots, \hat{b}_{(p)}$, and define the adjacent distances of the lasso estimators as $\hat{D}_1 = \hat{b}_{(1)}, \hat{D}_2 = \hat{b}_{(2)} - \hat{b}_{(1)}, \ldots, \hat{D}_p = \hat{b}_{(p)} - \hat{b}_{(p-1)}$.

Theorem 4.1. *Let* $i_\lambda = \min\{i : C^i_{under} \geq R, C^i_{upper} \geq \frac{1}{C}, D_i > (1-\frac{R}{C^i_{under}})^{-1}(1+R)\delta_\lambda\}$, *and* $S_\lambda = \{j : b^*_j \geq b^*_{(i_\lambda)}\}$. *Under the compatibility condition, if* $\lambda > 2\lambda_0$, *the following inequalities hold for the lasso estimator with a probability of at least* $1 - 2e^{-t^2/2}$:

$$\frac{\hat{D}(S_\lambda, S^c_\lambda)}{\hat{D}_{\max}(S^c_\lambda)} > R,$$
$$\frac{\hat{D}_{\max}(S_\lambda)}{\hat{D}(S_\lambda, S^c_\lambda)} \leq R,$$

where $\hat{D}(S_\lambda, S^c_\lambda) = \hat{D}_{p-s+i_\lambda}$, $\hat{D}_{\max}(S^c_\lambda) = \max\{\hat{D}_j : 1 \leq j < p - s + i_\lambda\}$ *and* $\hat{D}_{\max}(S_\lambda) = \max\{\hat{D}_j : p - s + i_\lambda < j \leq p\}$.

Let $\hat{S}_\lambda$ denote the important features identified at λ. Theorem 4.1 implies that $P(\hat{S}_\lambda = S_\lambda) \geq 1 - 2e^{-t^2/2}$. When the tuning parameter λ is bounded by $\min_{j\in S} |\beta^*_j| > (1 + R)4\lambda s/\phi^2$, we recover S exactly with a probability close to 1:

$$P(\hat{S}_\lambda = S) \geq 1 - 2e^{-t^2/2}.$$

Further, we conclude that for larger values of λ, i.e. such that $\max_{j\in S} |\beta^*_j| > (1 + R)4\lambda s/\phi^2 \geq \min_{j\in S} |\beta^*_j|$, there are no false positive signals in $\hat{S}_\lambda$, $P(\hat{S}_\lambda \subset S) \geq 1 - 2e^{-t^2/2}$. The following theorem then follows immediately from the fact that our SPSP estimator is $\hat{S} = \cup_\lambda \hat{S}_\lambda$:

Theorem 4.2. *Let* $i_{2\lambda_0} = \min\{i : C^i_{under} \geq R, C^i_{upper} \geq \frac{1}{C}, D_i > (1 - \frac{R}{C^i_{under}})^{-1}(1 + R)2\delta_0\}$ *and let* $S_{2\lambda_0} = \{j : b^*_j \geq b^*_{(i_{2\lambda_0})}\}$. *Under the compatibility condition, the SPSP estimator* $\hat{S}$ *over* $\lambda \in [2\lambda_0, \frac{\phi^2 D_{\max}}{4s(1+R)})$ *recovers* $S_{2\lambda_0}$ *with a probability of at least* $1 - 2e^{-t^2/2}$:

$$P(\hat{S} = S_{2\lambda_0}) \geq 1 - 2e^{-t^2/2}.$$

In particular, when $\min_{j\in S} |\beta^*_j| > (1 + R)2\delta_0$,

$$P(\hat{S} = S) \geq 1 - 2e^{-t^2/2}.$$

Theorem 4.2 suggests that the proposed SPSP procedure is consistent for variable selection under only the compatibility condition. With Theorem 4.2, we require that the tuning parameter λ is not larger than $\frac{\phi^2 D_{\max}}{4s(1+R)}$, the lower bound of which can be obtained with prior information. When no such information is available, we would need SPSP estimators over larger values of λ to not select any spurious variables. This is easy to verify in practice, since we can simply examine whether any new variables enter into the relevant set for larger values of λ. However, in order to theoretically guarantee such behavior of the solution paths, we need an additional condition, which is still substantially weaker than the irrepresentable condition:

Assumption 4.2. ***Identifiability Condition*** *Let $\eta > 0$ be some constant. For any $\bar{\boldsymbol{\beta}} = (\bar{\boldsymbol{\beta}}_S, \bar{\boldsymbol{\beta}}_{S^C})$, the following identifiability condition holds:*

$$\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\bar{\boldsymbol{\beta}}_S - \mathbf{X}_{S^C}\bar{\boldsymbol{\beta}}_{S^C}\|^2 \geq \min_{\|\boldsymbol{\beta}_S\|_1 \leq \|\bar{\boldsymbol{\beta}}_S\|_1 + (1-\eta)\|\bar{\boldsymbol{\beta}}_{S^c}\|_1} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\boldsymbol{\beta}_S\|^2. \quad (4.1)$$

The identifiability condition indicates that with the true set of relevant variables, we can approximate the noiseless response $\mathbf{X}\boldsymbol{\beta}^*$ at least as well as with any other set of variables under almost the same l_1 constraint. It is not difficult to verify that the condition is weaker than the irrepresentable condition.

Lemma 4.2. *The irrepresentable condition implies the identifiability condition.*

On the right hand side of (4.1), the coefficients of the irrelevant variables are set as $\boldsymbol{\beta}_{S^C} = 0$. In fact, we can further weaken the identifiability condition by allowing $\boldsymbol{\beta}_{S^c}$ to be non-zero on the right-hand side of (4.1). Instead, we only require $\|\boldsymbol{\beta}_{S^c}\|_1$ to be smaller than $\|\boldsymbol{\beta}_S\|_1$ up to a constant k. On top of that, we also relax the inequality (4.1) by taking $\kappa\eta\|\hat{\boldsymbol{\beta}}_{S^c}\|_1$ on the right side. As a result, we obtain the following weak identifiability condition:

Assumption 4.3. ***Weak Identifiability Condition*** *Let $\eta > 0$ be some constant. For any $\bar{\boldsymbol{\beta}} = (\bar{\boldsymbol{\beta}}_S, \bar{\boldsymbol{\beta}}_{S^C})$, then for $k = \frac{2}{2s+Rs(s+1)}$ and some κ that satisfies*

$$D_{\max} > \lambda_0 \frac{4s(1+R)}{\phi^2}\left\{\frac{Rs^2 + (2+R)S + 2}{\eta} - 1 + \kappa\right\},$$

the following weak identifiability condition holds,

$$\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\bar{\boldsymbol{\beta}}_S - \mathbf{X}_{S^C}\bar{\boldsymbol{\beta}}_{S^C}\|^2 \geq \min_{\boldsymbol{\beta}\in\Theta(\|\bar{\boldsymbol{\beta}}_S\|_1, \|\bar{\boldsymbol{\beta}}_{S^C}\|_1)} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\boldsymbol{\beta}\|^2 - \kappa\eta\|\bar{\boldsymbol{\beta}}_{S^c}\|_1, \quad (4.2)$$

where $\Theta(\|\bar{\boldsymbol{\beta}}_S\|_1, \|\bar{\boldsymbol{\beta}}_{S^C}\|_1) = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_S, \boldsymbol{\beta}_{S^C}) : \|\boldsymbol{\beta}\|_1 \leq \|\bar{\boldsymbol{\beta}}_S\|_1 + (1-\eta)\|\bar{\boldsymbol{\beta}}_{S^C}\|_1, \|\boldsymbol{\beta}_{S^C}\|_1 \leq k\|\boldsymbol{\beta}_S\|_1\}$.

Henceforth, we refer to the preceding weak identifiability condition with constants k and κ as $WIC(k,\kappa)$. Apparently, $WIC(0,0)$ is simply the identifiability condition in (4.1), and $WIC(k,\kappa)$ always implies $WIC(k',\kappa')$ for $k' > k, \kappa' > \kappa$. Therefore Assumption 4.3 is always weaker than Assumption 4.2. The above assumption ensures that when λ is large, the lasso estimates for the zero coefficients will not be much larger than those for the non-zero coefficients, so that we will not have any false positive signals from our SPSP procedure in $\hat{S}$. We combine this consideration with Theorem 4.1 to obtain the following result for the entire solution paths under the compatibility condition and $WIC(k,\kappa)$:

Theorem 4.3. *Suppose that under the compatibility condition and $WIC(k,\kappa)$ with $k = \frac{2}{2s+Rs(s+1)}$,*

$$D_{\max} > \lambda_0 \frac{4s(1+R)}{\phi^2}\left\{\frac{Rs^2 + (2+R)S + 2}{\eta} - 1 + \kappa\right\}.$$

Then the SPSP procedure over $\lambda \in [2\lambda_0, \infty)$ *identifies* $S_{2\lambda_0} = \{j : |\beta_j^*| > (1+R)2\delta_0\}$ *with a probability of at least* $1 - 2e^{-t^2/2}$, *i.e.*

$$P(\hat{S} = S_{2\lambda_0}) \geq 1 - 2e^{-t^2/2}.$$

When the true values of the coefficients are of a higher order than $\sqrt{\frac{\log p}{n}}$, it follows immediately from Theorem 4.3 that the asymptotic probability of identifying the true relevant set S is 1. Here we only need the tuning parameter λ to be not too small: $\lambda \geq 2\lambda_0$; unlike the existing literature, we do not require the tuning parameter λ to be in a specific region. In fact, we do not need consistent variable selection for any value of λ. We only need to control the bias of the lasso estimators for smaller λ's and to control the l_1 norm of the lasso estimators of those zero coefficients for larger λ's; both of these results require weaker conditions than achieving selection consistency at certain values of λ. By combining these results, the SPSP procedure can accomplish feature selection consistency under substantially weaker conditions, without a proper choice of the tuning parameter.

The theoretical results for other penalty functions can be developed following the paradigm as described above. We can derive Theorem 4.2 given the bias of the penalized estimators and the corresponding condition. For the non-convex penalties, the bias is smaller than lasso. For l_2 penalty, Shao and Deng (2012) gives the bias and convergence rate for high-dimensional ridge regression. To obtain Theorem 4.3, we need a similar identifiability condition as those in (4.1) and (4.2). However, the constraint on the l_1 norm needs to be replaced by the constraint on l_2 norm for ridge regression, and by the corresponding function form for the non-convex penalties.

5. Simulation studies

All of the following simulations are generated from the linear model (1.1) with $x_{ij} \sim N(0,1), i = 1, \ldots, n, j = 1, \ldots, p$, and $\epsilon_i \sim N(0, \sigma^2)$. The details of the simulation setups are as follows:

(M1) (Moderate Correlation, $p > n$) (Tibshirani, 1996) Let $\beta_1^* = 3, \beta_2^* = 1.5$, and $\beta_5^* = 2$, and let the remaining coefficients equal zero. The correlation between x_{j_1} and x_{j_2} is $0.5^{|j_1 - j_2|}$. We set $n = 50$, $p = 100$, and $\sigma = 3$.

(M2) (Moderate Correlation, $p >> n$) The setup here is the same as (M1), except that $p = 1000$.

(M3) (High Correlation, $p > n$) (Wang et al., 2011) Let $\boldsymbol{\beta}^* = (3, 3, -2, 3, 3, -2, 0, \ldots, 0)$, so that the first 6 coefficients are non-zero and the remaining 94 coefficients equal zero. The pairwise correlation between the first 3 variables is 0.9, the pairwise correlation between the second 3 variables is also 0.9, and the remaining 94 variables are independent of each other. Furthermore, the first 3 variables, the second 3 variables, and the remaining 94 variables are independent of each other. We set $n = 50$ and $\sigma = 3$.

(M4) (Misspecified Model, $p > n$) (Lv and Liu, 2014) The true values of the coefficients are $\boldsymbol{\beta}^* = (1, -1.25, 0.75, -0.95, 1.5, 0, \ldots, 0)$, where $p = 100$, the first 5 coefficients are non-zero, and the remaining coefficients equal zero. All of the variables are independent of each other, i.e the rows of $\mathbf{X}$ are generated from $N(\mathbf{0}, \mathbf{I}_p)$. We set $n = 50$ and $\sigma = 1$. Here the response $\mathbf{y}$ is generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{x}_{p+1} + \boldsymbol{\varepsilon}$, where $\mathbf{x}_{p+1} = \mathbf{x}_1 \circ \mathbf{x}_2$ is an interaction term between x_1 and x_2. However, x_{p+1} is not a candidate variable, and therefore the true model is not contained in the candidate models.

For all simulation examples, we conduct the SPSP procedure on the solution paths for lasso, adaptive lasso, SCAD, and MCP. One compelling advantage of the proposed SPSP procedure is that it can be applied for the l_2 penalty. Therefore, we also implement the SPSP algorithm with ridge regression for these examples. We compute the solution paths for convex penalties with the **R** package *glmnet* and those for SCAD and MCP with the **R** package *ncvreg*. All of the solution paths are computed over a grid of $K = 100$ values of the tuning parameter λ.

In Table 2–5, we compare the performance of the proposed SPSP procedure with the 10-fold cross-validation, generalized cross-validation, AIC, BIC, extended BIC and stability selection over 500 replicates for each setup. We record the following measures: FP, the number of false positives, FN, the number of false negatives, and ME, the Model Error $= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \hat{\boldsymbol{\Omega}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)/\sigma^2$, where $\hat{\boldsymbol{\Omega}}$ is the sample covariance of $\mathbf{X}$. We report the mean values of false positives and false negatives along with the median of the model errors, because the distribution of model error values is heavily skewed. We also report the standard errors for the above measures in parentheses.

Table 2 summarizes the results for M1, and we can observe that cross-validation, generalized cross-validation, AIC, and BIC on all the penalties tend to select too many variables in the model, as evidenced by the extremely large false positive numbers. The performance of the extended BIC here is comparable to that of SPSP, since SPSP produces a smaller number false negatives and slightly larger number of false positives. Stability selection performs well for adaptive lasso, but yields large false negative values for the other three penalties, especially the two non-convex ones: it misses two of the three signals on average for MCP and SCAD. The SPSP procedure provides the smallest model error for all penalties except for adaptive lasso, where the model error for SPSP is slightly larger than that for stability selection.

The simulation results for M2 are summarized in Table 3. We find that our SPSP approach has a more significant advantage over all of the other approaches when $p >> n$, compared to the results for M1. The SPSP procedure produces much smaller false positive values than cross-validation, generalized cross-validation, AIC, and BIC for lasso and adaptive lasso, while its false negative values are close to those for the other approaches. On the other hand, the extended BIC and stability selection often identify 1 or no important signals, although their false positive values are slightly smaller than the false positive value for SPSP. Here cross-validation has a similar performance to SPSP for the

non-convex penalties, but it performs rather poorly for lasso and adaptive lasso. Finally, the performance of SPSP with ridge regression is very close to that of the other penalties, which proves that even the l_2 penalty can be effective in feature selection with the SPSP procedure.

Table 4 presents the simulation results for M3, where the true model is very sparse and all of the important variables are correlated. We observe that compared to cross-validation, generalized cross-validation, AIC, and BIC, the SPSP procedure can dramatically improve the selection accuracy by selecting fewer irrelevant variables. We also observe that stability selection identifies fewer than one of the six signals with the non-convex penalties. Moreover, SPSP produces the smallest model error among all the approaches for all of the penalties, although in this case, the extended BIC also shows competitive performance.

The results for M4 are shown in Table 5. We observe that when models are misspecified, the SPSP procedure enjoys a substantially better performance for all four penalties in both selection accuracy and model estimation compared to the other selection criteria, except that the performance of cross-validation is similar to SPSP for the two non-convex penalties.

For each model, we further select one example to illustrate the advantage of the SPSP approach on the solution paths of lasso. We plot these solution paths and the partitions using SPSP in Figure 4.

In summary, the proposed SPSP approach provides the best or close-to-best performance for all of the penalties in all simulation examples. Generally speaking, cross-validation, generalized cross-validation, AIC, and BIC tend to select too many spurious variables, while the extended BIC and stability selection tend to ignore important features and in some cases even fail to identify any useful signals. Moreover, both stability selection and cross-validation cost significantly more computation time than SPSP. In our simulation studies, the computational cost of stability selection is 30–60 times more than that of SPSP. Finally, the performance of SPSP with ridge regression is quite similar to the other penalties, also generally outperforming the other selection criteria. This establishes the promise of potential applications of selection using the l_2 penalty, a unique feature of SPSP.

6. Data analysis

In this section, we apply the proposed SPSP procedure along with the other selection criteria to analyze the glioblastoma (GBM) gene expression data from The Cancer Genome Atlas (TCGA) consortium (*https://xenabrowser.net/datapages/*). The goal of the analysis is to identify the highly informative genes regarding the survival time of glioblastoma. The two gene expression data sets we obtained were both measured experimentally by the University of North Carolina TCGA genomic characterization center. In our analysis, we use the logarithm of the survival time as the response variable. The sample sizes of the two data sets are 428 and 91, after removing one common sample from the two data sets. We use the larger data set as our training set and the smaller one as the testing set.

TABLE 2
Simulation results for Model 1 over 100 replicates.

		CV	GCV	AIC	BIC	EBIC	STAB	SPSP
Lasso	FP	12.508	21.154	43.082	41.312	0.23	0.024	4.476
		(0.444)	(0.6)	(0.12)	(0.332)	(0.024)	(0.007)	(0.393)
	FN	0.042	0.064	0.088	0.092	0.848	1.166	0.37
		(0.009)	(0.012)	(0.013)	(0.014)	(0.047)	(0.031)	(0.027)
	ME	0.271	0.442	0.806	0.798	0.456	0.445	0.253
		(0.01)	(0.016)	(0.01)	(0.012)	(0.034)	(0.015)	(0.018)
adaLasso	FP	9.084	9.876	18.402	3.216	0.212	0.504	1.558
		(0.302)	(0.294)	(0.32)	(0.209)	(0.022)	(0.032)	(0.161)
	FN	0.036	0.042	0.032	0.104	0.628	0.382	0.446
		(0.009)	(0.009)	(0.008)	(0.015)	(0.035)	(0.024)	(0.03)
	ME	0.22	0.269	0.414	0.17	0.225	0.159	0.209
		(0.01)	(0.01)	(0.01)	(0.009)	(0.018)	(0.011)	(0.014)
SCAD	FP	2.388	18.496	19.476	17.736	1.414	0	2.22
		(0.103)	(0.132)	(0.125)	(0.187)	(0.195)	(0)	(0.186)
	FN	0.442	0.408	0.4	0.42	0.838	1.722	0.492
		(0.024)	(0.024)	(0.024)	(0.025)	(0.037)	(0.026)	(0.03)
	ME	0.28	0.8	0.816	0.787	0.35	0.66	0.251
		(0.011)	(0.013)	(0.013)	(0.014)	(0.028)	(0.026)	(0.015)
MCP	FP	0.928	17.902	18.192	17.692	2.972	0	2.38
		(0.06)	(0.124)	(0.121)	(0.134)	(0.286)	(0)	(0.167)
	FN	0.678	0.428	0.426	0.428	0.82	2.302	0.574
		(0.026)	(0.024)	(0.024)	(0.024)	(0.03)	(0.028)	(0.029)
	ME	0.268	0.873	0.873	0.873	0.265	0.773	0.266
		(0.01)	(0.012)	(0.012)	(0.012)	(0.019)	(0.049)	(0.015)
Ridge	FP							3.282
								(0.621)
	FN							0.932
								(0.038)
	ME							0.472
								(0.02)

We first screen the 17814 genes with the training data using sure independence screening (Fan and Lv, 2008) to identify the 1000 genes which are most correlated with the response (Wang et al., 2011). We then apply the proposed SPSP approach and the other methods to the training set and evaluate the prediction errors of the responses using the test data.

Table 6 shows the number of genes selected in the training set and the corresponding mean prediction error in the test set for each method. We notice that although the SPSP for lasso selects only 4 genes, it yields the smallest prediction error among all the penalty–criterion combinations. On the other hand, cross-validation, generalized cross-validation, AIC, and BIC for lasso all select too many irrelevant genes and therefore produce larger prediction errors. Both the extended BIC and stability selection select fewer variables and produce larger prediction errors. In fact, the extended BIC fails to recognize any signals for any of the penalties, and stability selection fails to recognize any signals for the non-convex penalties.

Two genes identified by the proposed SPSP approach for lasso have been proved to be important by previous studies. MAGEC2, the Melanoma-associated

TABLE 3
Simulation results for Model 2 over 100 replicates.

		CV	GCV	AIC	BIC	EBIC	STAB	SPSP
Lasso	FP	23.572	36.526	48.302	48.178	0.02	0.002	2.126
		(0.632)	(0.601)	(0.122)	(0.12)	(0.006)	(0.002)	(0.212)
	FN	0.142	0.13	0.122	0.122	2.25	2.3	0.712
		(0.017)	(0.016)	(0.015)	(0.015)	(0.043)	(0.025)	(0.033)
	ME	0.51	0.765	0.895	0.895	1.877	0.776	0.428
		(0.014)	(0.016)	(0.011)	(0.011)	(0.044)	(0.044)	(0.017)
adaLasso	FP	31.228	28.382	38.772	36.354	0.032	0.208	3.138
		(0.431)	(0.388)	(0.159)	(0.372)	(0.008)	(0.021)	(0.305)
	FN	0.126	0.138	0.13	0.132	1.602	0.956	0.672
		(0.016)	(0.016)	(0.016)	(0.016)	(0.045)	(0.033)	(0.033)
	ME	0.671	0.666	0.789	0.781	0.75	0.383	0.449
		(0.014)	(0.014)	(0.012)	(0.014)	(0.04)	(0.016)	(0.017)
SCAD	FP	6.086	22.248	23.842	22.302	0.084	0	6.53
		(0.213)	(0.135)	(0.154)	(0.143)	(0.013)	(0)	(0.399)
	FN	0.602	1.006	1.01	1.006	2.112	2.222	0.564
		(0.027)	(0.029)	(0.029)	(0.029)	(0.045)	(0.024)	(0.031)
	ME	0.389	0.887	0.895	0.887	1.469	0.751	0.559
		(0.013)	(0.012)	(0.012)	(0.012)	(0.049)	(0.041)	(0.021)
MCP	FP	1.384	15.378	15.574	15.266	3.786	0	3.408
		(0.079)	(0.072)	(0.074)	(0.075)	(0.269)	(0)	(0.214)
	FN	0.986	0.928	0.928	0.928	1.338	2.67	0.922
		(0.029)	(0.028)	(0.028)	(0.028)	(0.033)	(0.021)	(0.032)
	ME	0.41	0.932	0.932	0.932	0.705	2.143	0.575
		(0.013)	(0.012)	(0.012)	(0.012)	(0.027)	(0.046)	(0.019)
Ridge	FP							3.134
								(1.79)
	FN							0.944
								(0.031)
	ME							0.472
								(0.013)

antigen C2, a member of the MAGE gene family, is expressed in a wide variety of tumor types, including breast cancer (Song et al., 2016) and glioma (Sasaki et al., 2001). Song et al. (2016) further demonstrated that the gene expression levels of MAGEC2 are highly positively correlated with the phosphorylated STAT3, the signal transducer and activator of transcription 3, in human tumor tissues, and the STAT3 activation has been shown to affect multiple endogenous negative regulators such as PIAS3 (protein inhibitor of activated STAT3) in glioblastoma multiforme tumors. Therefore, the results of our present analysis provide additional evidence regarding the impact of MAGEC2 on glioblastoma behavior besides the work of Song et al. (2016), which may motivate scientists to conduct more experiments to explore the direct relationship between MAGEC2 and the human glioma. Moreover, the SPSP procedure additionally identifies TMEM49, also named VMP1, vacuole membrane protein 1, the gene which encodes an important protein in the process of autophagy. Ying et al. (2011) identified the gene as the direct and functional target of MicroRNA-210 (miR-210) and investigated the associations between the expression of the gene and hepatocellular carcinoma (HCC). Lai et al. (2015) demonstrated that miR-210

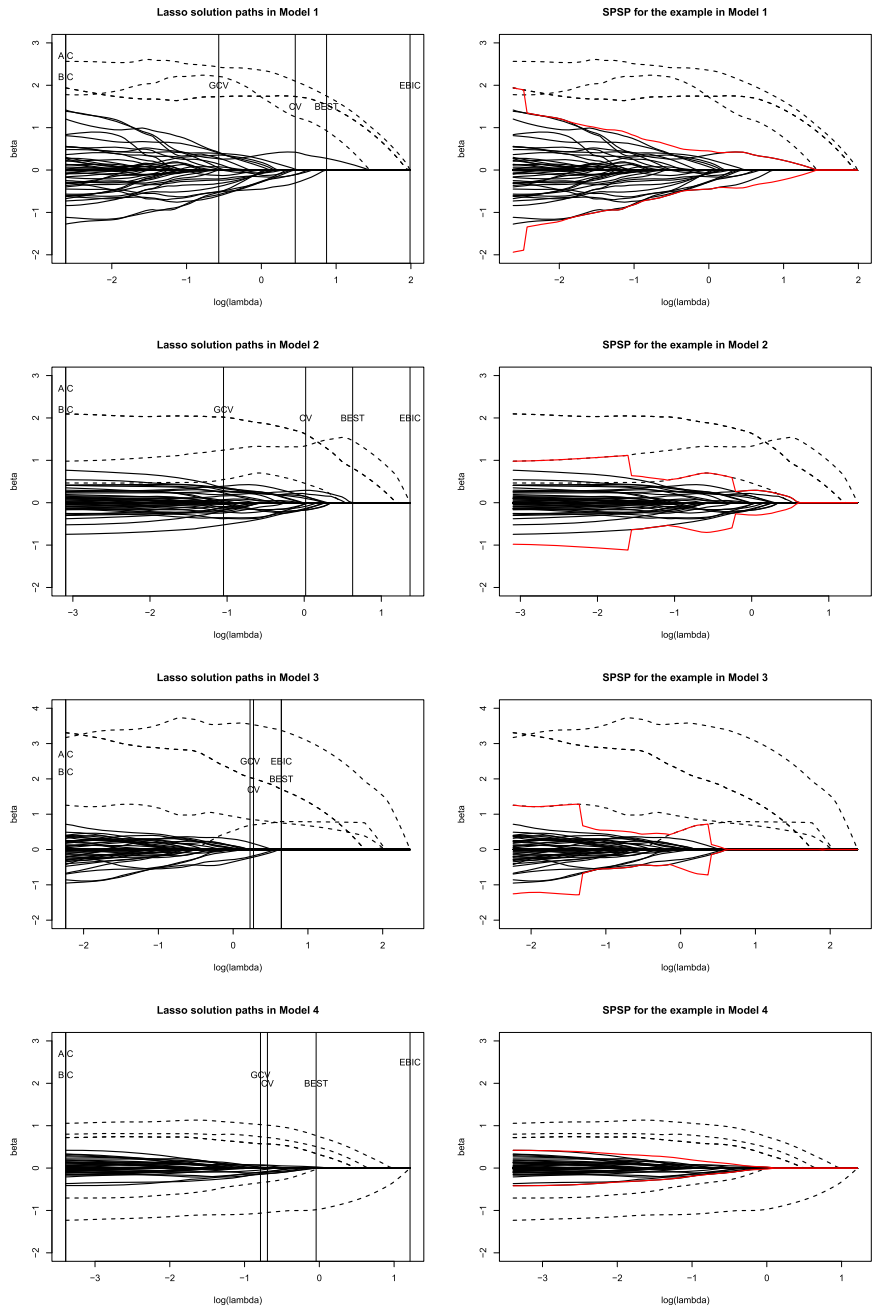

FIG 4. *Left Figures: The lasso solution paths for one simulated examples of these four models. The dashed lines are the paths of the non-zero coefficients, while the black lines are the paths of the zero coefficients The vertical lines represent the tuning parameters selected by different criteria. Right Figures: Partitions of the lasso solution paths of the same simulated examples using SPSP.*

Table 4
Simulation results for Model 3 over 100 replicates.

		CV	GCV	AIC	BIC	EBIC	STAB	SPSP
Lasso	FP	12.846	22.674	42.84	42.08	0.354	0.02	3.222
		(0.459)	(0.591)	(0.117)	(0.207)	(0.111)	(0.006)	(0.311)
	FN	2.27	2.268	2.208	2.224	3.124	4.754	2.6
		(0.026)	(0.028)	(0.032)	(0.031)	(0.056)	(0.035)	(0.038)
	ME	0.395	0.543	0.827	0.823	0.623	1.844	0.306
		(0.01)	(0.013)	(0.011)	(0.011)	(0.049)	(0.068)	(0.016)
adaLasso	FP	10.144	11.572	21.172	4.026	0.146	0.554	2.06
		(0.32)	(0.325)	(0.268)	(0.285)	(0.019)	(0.033)	(0.183)
	FN	1.83	1.794	1.218	2.274	2.92	3.536	2.424
		(0.036)	(0.037)	(0.039)	(0.034)	(0.038)	(0.032)	(0.043)
	ME	0.339	0.37	0.507	0.29	0.365	0.337	0.292
		(0.009)	(0.01)	(0.01)	(0.009)	(0.021)	(0.026)	(0.013)
SCAD	FP	1.308	19.776	20.894	18.28	0.594	0	0.696
		(0.092)	(0.147)	(0.139)	(0.265)	(0.11)	(0)	(0.126)
	FN	3.86	3.96	3.958	3.962	4.01	5.378	3.002
		(0.017)	(0.009)	(0.009)	(0.009)	(0.012)	(0.03)	(0.035)
	ME	0.39	0.813	0.827	0.804	0.37	2.207	0.267
		(0.009)	(0.012)	(0.012)	(0.014)	(0.019)	(0.067)	(0.009)
MCP	FP	0.574	18.718	19.002	18.506	2.438	0	0.984
		(0.058)	(0.127)	(0.125)	(0.135)	(0.29)	(0)	(0.131)
	FN	4	3.984	3.984	3.984	3.998	5.808	3.986
		(0)	(0.006)	(0.006)	(0.006)	(0.002)	(0.02)	(0.005)
	ME	0.375	0.906	0.906	0.906	0.363	3.572	0.353
		(0.007)	(0.012)	(0.012)	(0.012)	(0.014)	(0.052)	(0.01)
Ridge	FP							12.79
								(1.248)
	FN							1.36
								(0.073)
	ME							0.34
								(0.039)

is a potential serum biomarker in the diagnosis and prognosis of glioma, and it is also worth conducting further experiments to establish the connection between the expression of this gene and glioblastoma. Finally, our SPSP approach is the only method which can identify both MAGEC2 and THEM49 without selecting more than 100 genes.

7. Discussion

We have proposed a novel selection procedure for the penalized likelihood approach based on the entire solution paths. By utilizing estimators over all values of the tuning parameter, we can obtain better selection accuracy than the commonly used approach of selecting only one tuning parameter based on certain criteria. Moreover, the proposed SPSP procedure also achieves selection consistency under conditions that are substantially weaker than the irrepresentable condition, which is almost necessary under the framework currently in usage. Another advantage of SPSP is that we can now carry out selection with a strictly convex l_2 penalty. Although the present paper mainly focuses on feature selection for linear models, the SPSP procedure can easily be applied to

TABLE 5
Simulation results for Model 4 over 100 replicates.

		CV	GCV	AIC	BIC	EBIC	STAB	SPSP
Lasso	FP	17.438	22.624	42.604	40.322	2.2	0.002	4.312
		(0.416)	(0.512)	(0.118)	(0.352)	(0.396)	(0.002)	(0.365)
	FN	0.142	0.124	0.098	0.13	3.746	3.112	1
		(0.018)	(0.017)	(0.014)	(0.02)	(0.084)	(0.044)	(0.052)
	ME	0.926	1.145	1.742	1.725	5.377	2.797	1.146
		(0.024)	(0.028)	(0.022)	(0.026)	(0.125)	(0.067)	(0.048)
adaLasso	FP	13.15	13.94	25.526	9.976	0.362	0.496	2.91
		(0.316)	(0.327)	(0.28)	(0.43)	(0.038)	(0.032)	(0.242)
	FN	0.25	0.236	0.136	0.36	2.534	1.098	1.048
		(0.023)	(0.023)	(0.016)	(0.029)	(0.083)	(0.041)	(0.049)
	ME	0.872	0.936	1.213	0.922	2.081	0.906	1.059
		(0.023)	(0.023)	(0.022)	(0.027)	(0.113)	(0.038)	(0.045)
SCAD	FP	3.914	19.242	20.484	19.006	5.114	0	4.618
		(0.109)	(0.13)	(0.128)	(0.159)	(0.35)	(0)	(0.272)
	FN	0.314	0.324	0.324	0.326	1.962	4.094	0.574
		(0.028)	(0.028)	(0.028)	(0.028)	(0.096)	(0.034)	(0.04)
	ME	0.918	1.707	1.741	1.707	1.902	4.035	0.996
		(0.034)	(0.025)	(0.025)	(0.026)	(0.128)	(0.079)	(0.039)
MCP	FP	1.602	18.282	18.664	18.204	11.228	0	3.96
		(0.079)	(0.118)	(0.113)	(0.123)	(0.394)	(0)	(0.223)
	FN	0.562	0.33	0.326	0.33	0.646	4.63	0.548
		(0.038)	(0.027)	(0.027)	(0.027)	(0.047)	(0.025)	(0.04)
	ME	0.829	1.831	1.836	1.831	1.711	5.516	1.029
		(0.037)	(0.024)	(0.024)	(0.024)	(0.047)	(0.084)	(0.041)
Ridge	FP							3.528
								(0.628)
	FN							2.336
								(0.059)
	ME							2.093
								(0.073)

TABLE 6
Analysis of the glioblastoma data set

		CV	GCV	AIC	BIC	EBIC	STAB	SPSP
Lasso	NUM	37	157	275	2	0	3	4
	MPE	1.38	1.215	1.644	1.317	1.302	1.579	1.143
adaLasso	NUM	92	100	150	5	0	14	24
	MPE	1.266	1.216	1.175	1.422	1.302	1.529	1.363
SCAD	NUM	45	127	127	127	0	0	3
	MPE	1.299	3.005	3.005	3.005	1.302	1.302	1.28
MCP	NUM	23	113	113	76	0	0	13
	MPE	1.277	2.472	2.472	1.882	1.302	1.302	1.502

most selection problems with one or more tuning parameters. In Appendix A.3, we include a simulation study on estimating high-dimensional Gaussian graphical models with the SPSP approach to illustrate SPSP's potential for other selection problems.

With this study, we hope to initiate a discussion of how to better apply information contained in entire solution paths. For example, we can rank the importance of features by exploring the differences between the behaviors of the solution paths for important features and those for spurious ones. It is

also possible to develop an inference procedure and quantify the uncertainty of selection results based on the entire solution paths. In addition, it might be interesting to see whether the solution paths for important features differ from those for irrelevant ones in any other manner besides the magnitude of the estimators.

8. Code

All source code for the simulations and real data analysis can be found on GitHub: https://github.com/yliu433/r-spsp.

Appendix A: Appendix

A.1. Technical proofs

Proof of Theorem 4.1. By Lemma 4.1, with probability at least $1-2e^{-t^2/2}$, we have

$$\hat{D}(S_\lambda^c) \leq \max\{D_{i'} : i' < i_\lambda\} + \delta_\lambda$$

and similarly for any $j_1 \in S_\lambda, j_2 \in S_\lambda^c$,

$$\begin{aligned}\left||\hat{\beta}_{j_1}| - |\hat{\beta}_{j_2}|\right| &\geq |\beta^*_{j_1}| - \left||\hat{\beta}_{j_1}| - |\beta^*_{j_1}|\right| - |\hat{\beta}_{j_2}| \\ &\geq D_{i_\lambda} - \delta_\lambda\end{aligned}$$

Then with probability at least $1-2e^{-t^2/2}$,

$$\frac{\hat{D}(S_\lambda, S_\lambda^c)}{\hat{D}(S_\lambda^c)} \geq \frac{D_{i_\lambda} - 2\delta_\lambda}{\max\{D_{i'} : i' < i_\lambda\} + \delta_\lambda} > R.$$

The above inequality follows immediately from $C_{under}^{i_\lambda} > R$ and $D_{i_\lambda} > (1 - \frac{R}{C_{under}^{i_\lambda}})^{-1}(1+R)\delta_\lambda$.

To verify the partitioning rule, we have

$$\hat{D}(S_\lambda) \leq D_{\max} + \delta_\lambda$$

Then with probability at least $1-2e^{-t^2/2}$,

$$\begin{aligned}\frac{\hat{D}(S_\lambda)}{\hat{D}(S_\lambda, S_\lambda^c)} &\leq \frac{D_{\max} + \delta_\lambda}{D_{i_\lambda} - \delta_\lambda}, \\ &\leq C + \frac{(C+1)\delta_\lambda}{D_{i_\lambda} - 2\delta_\lambda} \\ &\leq C + \frac{1+C}{R} = R.\end{aligned}$$

The last inequality follows from the fact that $R = 1 + C$. □

Proof of Theorem 4.2. A sufficient condition for $\hat{S}_\lambda \cap S^C = \emptyset$ is

$$D_{\max} > R\delta_\lambda.$$

Then the theorem follows immediately from Theorem 4.1 and the fact that $\hat{S} = \cup_\lambda \hat{S}_\lambda$. □

Proof of Lemma 4.2. Note that

$$\begin{aligned} & \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\hat{\boldsymbol{\beta}}_S - \mathbf{X}_{S^C}\hat{\boldsymbol{\beta}}_{S^C}\|^2 \\ \geq \quad & \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\hat{\boldsymbol{\beta}}_S - \mathbf{X}_S \mathrm{diag}(\mathrm{sign}(\boldsymbol{\beta}^*_S))\,\mathrm{sign}(\boldsymbol{\beta}^*_S)(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\mathbf{X}_{S^C}\hat{\boldsymbol{\beta}}_{S^C}\|^2. \end{aligned}$$

By the irrepresentable condition, there is a $\eta > 0$ such that

$$\|\,\mathrm{sign}(\boldsymbol{\beta}^*_S)(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\mathbf{X}_{S^C}\|_\infty \leq 1-\eta,$$

from which it follows immediately that

$$\|\mathrm{diag}(\mathrm{sign}(\boldsymbol{\beta}^*_S))\,\mathrm{sign}(\boldsymbol{\beta}^*_S)(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\mathbf{X}_{S^C}\hat{\boldsymbol{\beta}}_{S^C}\|_1 \leq \|\hat{\boldsymbol{\beta}}_{S^C}\|_1(1-\eta).$$

Therefore, the identifiability condition holds if the irrepresentable condition holds.

□

To prove Theorem 4.3, we introduce the following two lemmas.

Lemma A.1. *Under $WIC(k,\kappa)$, the following inequality holds for the lasso solution $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\beta}}_{S^C})$ with $\lambda > \lambda_0(\frac{2+2k-k\eta}{k\eta} + \kappa)$ with probability at least $1-2e^{-t^2/2}$,*

$$\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 \leq k\|\hat{\boldsymbol{\beta}}_S\|_1.$$

Proof. Since $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\beta}}_{S^C})$ is the lasso solution, then for

$$\tilde{\boldsymbol{\beta}}_S = \arg \min_{\|\boldsymbol{\beta}_S\|_1 \leq \|\hat{\boldsymbol{\beta}}_S\|_1 + (1-\eta)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\boldsymbol{\beta}_S\|^2,$$

we have

$$\begin{aligned} & \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\tilde{\boldsymbol{\beta}}_S\| + \lambda\|\tilde{\boldsymbol{\beta}}_S\|_1 + \frac{1}{n}2\varepsilon^T\mathbf{X}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \\ \geq & \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\hat{\boldsymbol{\beta}}_S - \mathbf{X}_{S^C}\hat{\boldsymbol{\beta}}_{S^C}\| + \lambda\|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{n}2\varepsilon^T\mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}). \end{aligned}$$

If the following inequality holds when $\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 > k\|\hat{\boldsymbol{\beta}}_S\|_1$

$$\lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\tilde{\boldsymbol{\beta}}_S\|_1 + \frac{1}{n}2\varepsilon^T\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \geq 0.$$

Then it follows from the identifiability condition that

$$\begin{aligned} & \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\tilde{\boldsymbol{\beta}}_S\| + \lambda\|\tilde{\boldsymbol{\beta}}_S\|_1 + 2\varepsilon^T\mathbf{X}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \\ \leq & \frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_S\hat{\boldsymbol{\beta}}_S - \mathbf{X}_{S^C}\hat{\boldsymbol{\beta}}_{S^C}\| + \lambda\|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{n}2\varepsilon^T\mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}), \end{aligned}$$

therefore either $\hat{\boldsymbol{\beta}}_{S^c} = 0$ or $\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 \leq k\|\hat{\boldsymbol{\beta}}_S\|_1$.

Because we have concluded $P(\|\frac{1}{n}2\varepsilon^T\mathbf{X}\|_\infty < \lambda_0) > 1 - 2e^{-t^2/2}$, when $\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 > k\|\hat{\boldsymbol{\beta}}_S\|_1$, we have

$$\begin{aligned}
& \lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\tilde{\boldsymbol{\beta}}_S\|_1 + \frac{1}{n}2\varepsilon^T\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \\
\geq\ & \lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\tilde{\boldsymbol{\beta}}_S\|_1 - \lambda_0\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_1 \\
\geq\ & \lambda\eta\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 - \lambda_0\|\tilde{\boldsymbol{\beta}}_S\|_1 - \lambda_0\|\hat{\boldsymbol{\beta}}\|_1 \\
\geq\ & \lambda\eta\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 - \lambda_0(\frac{1}{k} + 1 - \eta)\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 - \lambda_0(\frac{1}{k} + 1)\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 \\
=\ & \{\lambda\eta - \lambda_0(\frac{2}{k} + 2 - \eta)\}\|\hat{\boldsymbol{\beta}}_{S^C}\|_1 \\
>\ & 0.
\end{aligned}$$

The last inequality follows from $\lambda > \lambda_0\frac{2+2k-k\eta}{k\eta}$. □

Lemma A.2. *Under the weak identifiability condition with* $k = \frac{2}{2s+Rs(s+1)}$,

$$P(\hat{S}_\lambda \subset S) \geq 1 - 2e^{-t^2/2},$$

for $\lambda > \lambda_0(\frac{2+2k-k\eta}{k\eta} + \kappa)$.

Proof. Denote $\hat{\beta}^{\max}_{S^C} = \max\{|\hat{\beta}_j| : j \in S^C\}$, and sort the absolute values in $\{|\hat{\beta}_j| : j \in S, |\hat{\beta}_j| \geq \hat{\beta}^{\max}_{S^C}\}$ in ascending order to get $\hat{\beta}^u_{(1)} \leq \hat{\beta}^u_{(2)} \cdots \leq \hat{\beta}^u_{(d)}$, where $d \leq s$ is the cardinality of the set $\{|\hat{\beta}_j| : j \in S, |\hat{\beta}_j| \geq \hat{\beta}^{\max}_{S^C}\}$. Let $\Delta_1 = \hat{\beta}^u_{(1)} - \hat{\beta}^{\max}_{S^C}$ and $\Delta_i = \hat{\beta}^u_{(i)} - \hat{\beta}^u_{(i-1)}$, $i = 2, \ldots, d$, then $\|\hat{\beta}_S\|_1 \leq s\hat{\beta}^{\max}_{S^C} + \sum_{i=1}^d i\Delta_i$. Therefore by Lemma A.1

$$\hat{\beta}^{\max}_{S^C} \leq \|\hat{\beta}_{S^C}\|_1 \leq ks\hat{\beta}^{\max}_{S^C} + k\sum_{i=1}^d i\Delta_i \leq ks\hat{\beta}^{\max}_{S^C} + k\frac{s(s+1)}{2}\Delta_{\max},$$

where $\Delta_{\max}$ is the maximum value of $\Delta_1, \ldots, \Delta_d$. It follows when $k = \frac{2}{2s+Rs(s+1)}$ that

$$\hat{\beta}^{\max}_{S^C} \leq \frac{ks(s+1)}{2(1-ks)}\Delta_{\max} = \frac{1}{R}\Delta_{\max}.$$ □

Proof of Theorem 4.3. The theorem follows directly from Theorem 4.2 and Lemma A.2. □

A.2. Simulation studies on the sensitivity of SPSP to the value of R

The constant R in the proposed SPSP algorithm controls the order of the magnitude in the partitioning rule. Here we examine the sensitivity of the SPSP procedure with respect to the choice of the constant R. Note that in the numerical studies we conducted, these constants are data adaptive rather than given arbitrarily.

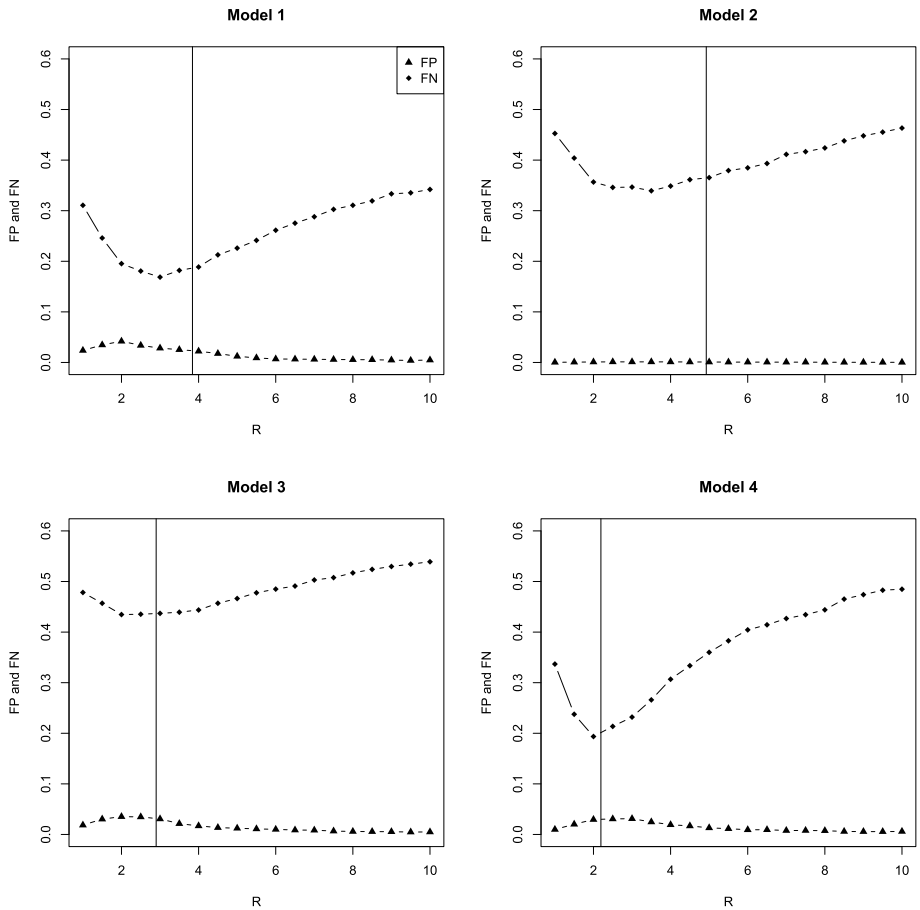

FIG 5. *The mean of the FPR and FNR over* 500 *replicates over different choices of the constant R in SPSP on the Lasso. The vertical lines in the graphs are the average values of the R estimated from the data set.*

In particular, we select a sequence of the numbers from 1 to 10 with the increment 0.5, as the candidates of the constant R. We use each number as the constant R in the proposed SPSP algorithm on lasso for all the models, and record the means of the false positive rates ($FPR = FP/\#$ of true zero features) and the false negative rates ($FNR = FN/\#$ of true nonzero features) for each value of R.

The results are shown in Figure 5. We observe that the means of the FPR and FNR are relatively stable across different choices of the constant R. Note that the vertical line in each graph represents the mean of the values of R estimated from the data set. It illustrates that the selection results of the SPSP algorithm are not so sensitive to the choice of the constant R as long as R stays within a reasonable range.

A.3. Simulation study on the Gaussian graphical model

Besides feature selection for linear models, the SPSP procedure introduced in the paper can be widely applied for other selection problems under the framework of penalized likelihood estimation. Here we simply present a simulation study to illustrate the performance of the SPSP algorithm for the Gaussian graphical model.

The data is simulated from $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, where the inverse of the covariance matrix is set as $(\mathbf{\Sigma}^{-11})_{j,j} = 1, (\mathbf{\Sigma}^{-11})_{j,j+1} = 0.5, (\mathbf{\Sigma}^{-11})_{j+1,j} = 0.5, j = 1, \cdots, p/4$, and zero otherwise. We set $p = 100$ and $n = 50$ in the simulation. This example is a AR(1) model, which has been used by Friedman, Hastie and Tibshirani (2008) and Yuan and Lin (2007) for the numerical study of the graphical lasso.

We compare the performance of the proposed SPSP algorithm with the graphical models selected by BIC and the EBIC. Here the details about using BIC and the EBIC to choose the tuning parameter in the graphical models are described in Foygel and Drton (2010). We apply the **R** package *glasso* to solve the graphical lasso estimators and apply the package *qgraph* to select the graphical lasso models by BIC and the EBIC. Note that the grid of the tuning parameters in the simulation is generated automatically by the function *glassopath*.

Table 7
The mean of FP, FN values of the SPSP algorithm, BIC, and the EBIC over 100 *replicates (Standard Error in the parentheses). The true model has* 25 *nonzero dependencies and* 4925 *zero dependencies.*

	SPSP	BIC	EBIC
FP	19.31	116.56	0
	(2.48)	(3.2)	(0)
FN	2.50	0	25
	(0.80)	(0.0)	(0)

We report the mean and the standard error of the number of the false positives (FP), the number of the false negatives (FN) of the SPSP algorithm, BIC and the EBIC over 100 replicates in Table 7. We observe that the BIC tends to include too many zero dependencies (high FP value) while the EBIC missed all the nonzero dependencies (high FN value) in the model. Compared with the results of these two criteria, the SPSP algorithm has a much better performance in terms of selection accuracy, which selects most of the nonzero dependencies without adding many zero dependencies in the model.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information, BN Petrow, F.* MR0483125

Atkinson, A. C. (1980). A Note on the Generalized Information Criterion for Choice of a Model. *Biometrika* **67** 413–418.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics* **32** 870–897. MR2065192

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705–1732. MR2533469

Bogdan, M., Ghosh, J. K. and Doerge, R. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167** 989–999.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 641–656. MR1979381

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer. MR2807761

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1** 169–194. MR2312149

Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *Information Theory, IEEE Transactions on* **51** 4203–4215. MR2243152

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

Cho, H. and Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: series B (statistical methodology)* **74** 593–622. MR2925375

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911. MR2530322

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961. MR2065194

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 531–552. MR3065478

Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. In *Advances in Neural Information Processing Systems*.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science* **14** 382–401. MR1765176

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of statistics* **38** 2282–2313. MR2676890

Lai, N., Wu, D., Fang, X., Lin, Y., Chen, S., Li, Z. and Xu, S. (2015).

Serum microRNA-210 as a potential noninvasive biomarker for the diagnosis and prognosis of glioma. *British journal of cancer* **112** 1241–1246.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 141–167. MR3153937

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462. MR2278363

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 417–473. MR2758523

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology* **53** 793–808.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92** 179–191. MR1436107

Sasaki, M., Nakahira, K., Kawano, Y., Katakura, H., Yoshimine, T., Shimizu, K., Kim, S. U. and Ikenaka, K. (2001). MAGE-E1, a new member of the melanoma-associated antigen gene family and its expression in human glioma. *Cancer research* **61** 4809–4814.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6** 461–464. MR0468014

Shao, J. and Deng, X. (2012). Estimation in High-Dimensional Linear Models with Deterministic Design Matrices. *The Annals of Statistics* **40** 812–831. MR2933667

Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232. MR2949354

Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika* **91** 785–800. MR2126033

Song, X., Hao, J., Wang, J., Guo, C., Wang, Y., He, Q., Tang, H., Qin, X., Li, Y., Zhang, Y. et al. (2016). The cancer/testis antigen MAGEC2 promotes amoeboid invasion of tumor cells by enhancing STAT3 signaling. *Oncogene*.

Sun, W., Wang, J. and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research* **14** 3419–3440. MR3144467

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108. MR2136641

van de Geer, S. (2007). The deterministic lasso. In *Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich*.

Wang, H., Li, G. and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 63–78. MR2301500

Wang, S., Nan, B., Rosset, S. and Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics* **5** 468–485. MR2810406

Ying, Q., Liang, L., Guo, W., Zha, R., Tian, Q., Huang, S., Yao, J., Ding, J., Bao, M., Ge, C. et al. (2011). Hypoxia-inducible MicroRNA-210 augments the metastatic potential of tumor cells by targeting vacuole membrane protein 1 in hepatocellular carcinoma. *Hepatology* **54** 2064–2075.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105** 312–323. MR2656055

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7** 2541–2563. MR2274449

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101** 1418–1429. MR2279469