JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 524, 1797–1812, Theory and Methods
https://doi.org/10.1080/01621459.2017.1377080

Sparsity Oriented Importance Learning for High-Dimensional Linear Regression

Chenglong Ye[a], Yi Yang[b], and Yuhong Yang[a]

[a]School of Statistics, University of Minnesota, Minneapolis, MN; [b]Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada

ABSTRACT

With now well-recognized nonnegligible model selection uncertainty, data analysts should no longer be satisfied with the output of a single final model from a model selection process, regardless of its sophistication. To improve reliability and reproducibility in model choice, one constructive approach is to make good use of a sound variable importance measure. Although interesting importance measures are available and increasingly used in data analysis, little theoretical justification has been done. In this article, we propose a new variable importance measure, sparsity oriented importance learning (SOIL), for high-dimensional regression from a sparse linear modeling perspective by taking into account the variable selection uncertainty via the use of a sensible model weighting. The SOIL method is theoretically shown to have the inclusion/exclusion property: When the model weights are properly around the true model, the SOIL importance can well separate the variables in the true model from the rest. In particular, even if the signal is weak, SOIL rarely gives variables not in the true model significantly higher important values than those in the true model. Extensive simulations in several illustrative settings and real-data examples with guided simulations show desirable properties of the SOIL importance in contrast to other importance measures. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2016
Revised August 2017

KEYWORDS
Adaptive regression by mixing; Model averaging; Reliability and reproducibility; Variable importance

1. Introduction

Variable importance has been an interesting research topic that helps to identify which variables are most important for understanding, interpretation, estimation, or prediction purposes. The potential usages of variable importance measures include: (1) they help reduce the list of variables to be considered by screening out those with importance values below a threshold. This leads to cost and time saving in data analysis; (2) they also help decision makers to obtain a more comprehensive understanding of the underlying data-generation process than trusting any single model by a variable selection procedure; (3) they offer a ranking of variables that can be used to consider model selection or model averaging in a nested fashion, which simplifies the consideration of all subset models; (4) they can help decision makers to change or replace variables based on practical considerations. See Feldman (2005), Louppe et al. (2013), Braun and Oswald (2011), Grömping (2015), Hapfelmeier et al. (2014), Archer and Kimes (2008), and Strobl et al. (2007) for reference.

Under the linear regression setting, various methods have been proposed for evaluating variable importance. The first type includes simple measures based on a final selected model, for example, t-test values, (standardized) regression coefficients, and p-values of the variables. This approach has the severe drawback associated with any "winner takes all" variable selection method. The variable selection uncertainty is totally ignored and all the non-selected variables have zero importance.

Another approach is based on the R^2 decomposition. Lindeman, Merenda and Gold (1980) used the improved explained variance averaged over all possible orderings of predictors to provide a ranking of the predictors. Feldman (2000) extended it to the weighted version (PMVD). Several encouraging methods, such as dominance analysis (Budescu 1993), hierarchical partitioning (Chevan and Sutherland 1991), information criterion based method (Theil and Chung 1988) and the product of standardized true coefficients and partial correlation (Hoffman 1960), have also been proposed.

Besides importance measuring with parametric models, nonparametric approaches are also available. For regression and classification, random forest (Breiman 2001) and its variants have attracted a lot of attention in many fields. Breiman (2001) proposed two versions of variable importances for random forest. Ishwaran (2007) studied the theoretical properties of variable importance for binary regression with random forest. There, the variable importance is defined as the difference between the prediction error before and after the variable is noised up. Under proper assumptions, the variable importance is shown to converge and suitably upper-bounded. Strobl et al. (2008) proposed conditional variable importance for random forest to correct the bias of variable importance when there exist correlated variables. Ferrari and Yang (2015) assessed variable importance from a variable selection confidence set (VSCS) perspective.

In this article, we propose a sparsity oriented importance learning (SOIL) for high-dimensional regression data. For our approach, by assigning weights to the candidate linear models (or generalized linear models for classification), we come up

CONTACT Yi Yang yi.yang6@mcgill.ca Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

with measures of importance of the predictors in an absolute scale in [0, 1].

Several features/advantages of our method can be concluded as follows. First, it involves multiple high-dimensional variable selection methods and combines all their solution path models, which produces many candidate models rather than being based on only one model selection method. The resulting importance values are thus more reliable than trusting one method alone. Second, SOIL uses external weighting, which is independent of the model selection methods. This can avoid possible bias brought up by using a method both for coming up with candidate models and for assessing the models for weighting. Third, from the main theorem in the article, we gain a theoretical understanding of our method. We prove that the importances of the true variables will tend to 1 and the importances of the other variables will tend to 0 as the sample size increases, as long as the weighting is sensible. Last but not least, compared with other importance measures, our method also shows excellent performances in the numerical study, with desirable behaviors such as *exclusion, inclusion, order preserving, robustness,* etc.

In the current era of rich high-dimensional data, with the well-recognized severe problem of irreproducibility of scientific findings (see, e.g., Ioannidis and Khoury 2011; McNutt 2014; Stodden 2015), we believe the use of informative importance measures can much improve the reliability of data analysis in multiple ways:

1. First, if the data analyst has already chosen a set of covariates for finalizing a model to be recommended, the SOIL importance measure is helpful to put the model under a more objective light. He/she can immediately inspect if some variables deemed important by SOIL are missing in the set or the other way around. If so, the analyst may want to investigate on the matter. For instance, residuals from the model based on the current set of covariates, when plotted against the missing variables, may reveal their relevance. Models with/without the variables in questions can be fit and compared for a better understanding on their usefulness.
2. Based on the theoretical properties of the SOIL, variables most suitable for sparse modeling receive higher importance values. Thus, the SOIL can be naturally used to find the best model for the data. In theory, any fixed cutoff in (0, 1) leads to a good performance (see Theorem 2). But the best cutoff depends on the purpose of the final model: for prediction accuracy, the cutoff should be lower and for identifying variables than can be validated at similar sample sizes in future studies, the cutoff should be higher. See, e.g., Yang (2005) to understand the subtle matter of the conflict between model identification and estimation/prediction.
3. Whether one comes up with a set of covariates based on SOIL importance (as described above) or not (e.g., using a penalized likelihood based model selection method), the SOIL importance values of the variables help the data analyst get a sense on model selection uncertainty. More specifically, if there are quite a few variables having importance values similar to some in a final model (obtained from a trustworthy process that has, at least reasonably, justified the usefulness of the selected covariates, e.g., based on cross-validation), it may indicate that the model selection uncertainty is perhaps high for the data and there are alternative choices of variables that can give similar predictive performances. In such a case, it is advantageous for the data analyst and the decision maker to be well-informed on possible alternative models/covariates to be used. For instance, if some covariates are much less costly for future experiments or operations, they may be preferred to be included in the final model even if their importance values are slightly lower than some other ones in a good model.
4. When estimating the regression function or prediction is the main goal, the understanding on degree of model selection uncertainty, together with other model selection diagnostic tools (see, e.g., Nan and Yang 2014 for references), can help the data analyst decide on the choice between model selection and model averaging (see Yang 2003; Chen, Giannakouros, and Yang 2007 for results on comparison between model selection and model averaging).

In summary, the SOIL method is helpful in different stages of model building. It can be used to narrow down the set of covariates for further consideration and for reaching a final model with sound considerations. Equally or even more importantly, it provides an objective view on reliability of the model and the model selection uncertainty. This gives information unavailable in the traditional practice of glorifying the final model and thus can help much improve reproducibility of data analysis that involves variable selection.

The remainder of the article is organized as follows. In Section 2, we introduce the proposed SOIL methodology and provide a theoretical understanding on some key aspects. Sections 3 and 4 present the details of choosing the candidate models and the weighting for SOIL in practice. In Section 5, we conduct several simulations that fairly and informatively compare the performance of SOIL and three existing and commonly used variable importance measures (LMG and two versions of random forest importances). Furthermore, we apply these methods to three real datasets in Section 6. A discussion about variable importance is then presented in Section 7, followed by the proofs of the results in the Appendix.

2. General Methodology

In this section, we introduce the *sparsity oriented importance learning* (SOIL) procedure, which provides an objective and informative profile of variable importances for high-dimensional regression and classification models. We consider the regression setting first, and the generalization to the classification model will be discussed later in Section 4.

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be the $n \times p$ design matrix with $X_j = (x_{1j}, \ldots, x_{nj})^\intercal$, $j = 1, \ldots, p$, and $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ be the n-dimensional response vector. The design matrix can also be written as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\intercal$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$, $i = 1, \ldots, n$. We consider the following underlying linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the vector of n independent errors and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\intercal$ is a p-dimensional vector of the true underlying model that generates the data. In general, predictors

may include those created by the original predictors observed, such as $\sqrt{X_1}$, X_1^2 and X_1X_3. We adopt the sparsity assumption that most regression coefficients β_j^* are zero. Denote by $|\cdot|$ the cardinality of a set. We assume $\boldsymbol{\beta}^*$ is r^*-sparse, where $r^* = |\mathcal{A}^*|$ with $\mathcal{A}^* \equiv \text{supp}(\boldsymbol{\beta}^*) = \{j : \beta_j^* \neq 0\}$.

SOIL importance depends on two ingredients: a manageable set of models (often based on a preliminary analysis) and a reliable external weighting method on the models. Together they can provide valuable information on importance of the predictors.

Suppose that one can obtain a collection of models $\boldsymbol{A} = \{\mathcal{A}_k\}_{k=1}^K$, which can be either a full list of all-subset models when p is small, or a group of models obtained from high-dimensional variable selection procedures such as Lasso (Tibshirani 1996), Adaptive Lasso (Zou 2006), SCAD (Fan and Li 2001) and MCP (Zhang 2010) etc., when p is large. We refer to $\mathcal{A}_k$, $k = 1, \ldots, K$ as *candidate models*, and $\mathbf{w} = (w_1, \ldots, w_K)^\intercal$ as the corresponding weighting vector, which is estimated from the data.

Given the set $\boldsymbol{A}$ and the weighting $\mathbf{w}$, we define the SOIL importance measure for the jth variable, $j \in \{1, \ldots, p\}$, as the accumulated sum of weights of the candidate models $\mathcal{A}^k$ that contains the jth variable. That is

$$\text{SOIL Importance}: \quad S_j \equiv S(j; \mathbf{w}, \boldsymbol{A}) = \sum_{k=1}^K w_k I(j \in \mathcal{A}^k).$$

2.1. Theoretical Properties

We will show consistency of the SOIL importance measure, under the condition that the weighting vector $\mathbf{w} = (w_1, \ldots, w_K)^\intercal$ satisfies the following properties referred to as *weak consistency* and *consistency:*

Definition 1 (Weak Consistency and Consistency). The weighting vector $\mathbf{w}$ is *weakly consistent* if

$$\frac{\sum_{k=1}^K w_k |\mathcal{A}^k \nabla \mathcal{A}^*|}{r^*} \xrightarrow{p} 0, \qquad \text{as} \quad n \to \infty, \tag{1}$$

and $\mathbf{w}$ is *consistent* if

$$\sum_{k=1}^K w_k |\mathcal{A}^k \nabla \mathcal{A}^*| \xrightarrow{p} 0, \qquad \text{as} \quad n \to \infty,$$

where ∇ denotes the symmetric difference of two sets and $|\cdot|$ denotes number counting.

Remark 1. Intuitively, both weak consistency and consistency of weighting ensure that the weighting of the candidate models is concentrated enough around the true model, but to different degrees. Including the denominator r^* in (1) makes the weak consistency condition more likely to be satisfied than consistency, when the true model size r^* is allowed to increase in dimension as n increases, as long as it satisfies the sparsity assumption $r^* << n$.

Remark 2. For a very poor candidate set $\boldsymbol{A}$, there may not exist any (weakly) consistent weighting vector.

Definition 2 (Path-consistent). A method is called path-consistent if

$$P(\mathcal{A}^* \in \Delta) \to 1, \text{ as } n \to \infty,$$

where Δ denotes the whole solution path produced by the method.

Remark 3. The definition of path-consistency provides an option of obtaining a good candidate set $\boldsymbol{A}$. We can consider the solution paths of multiple path-consistent methods, which will be further discussed in Section 3.1.

There are several different methods in the literature for providing the weight vector $\mathbf{w} = (w_1, \ldots, w_K)^\intercal$ for the candidate models $\boldsymbol{A}$. For example, Buckland, Burnham, and Augustin (1997) and Leung and Barron (2006) studied a weighting method based on information criterion, such as AIC (Akaike 1973) and BIC (Schwarz et al. 1978); Hoeting et al. (1999) proposed the weighting by Bayesian model averaging (BMA) from a Bayesian perspective; several attractive frequentist model averaging approaches are also developed (e.g., Yang 2001; Hjort and Claeskens 2003; Buckland, Burnham, and Augustin 1997; Hansen 2007; Liang et al. 2011; Cheng, Ing, and Yu 2015; Cheng and Hansen 2015). In particular, Yang (2001) proposed a weighting strategy by data splitting and cross-assessment, which is referred to as the adaptive regression by mixing (ARM). He proved that the weighting by ARM delivers the best rate of convergence for regression estimation. One advantage of ARM is that it can be applied to combine general regression procedures (not limited to parametric models). The ARM weighting was extended to the classification problems in Yang (2000), Yuan and Ghosh (2008), and Zhang, Lu, and Zou (2013).

Among the aforementioned weighting methods, there are several that give consistent weights $\mathbf{w}$. For example, when there are a fixed number of models in the candidate model set, BMA typically gives a consistent weighting. ARM also gives consistent weighting when the data splitting ratio is properly chosen (Yang 2007). Now, we prove that (a) under the assumption of weakly consistent weighting, the sum of the SOIL importance of the true variables will tend to the size of the true model r^*, while the sum of the SOIL importance of the variables excluded by the true model converges to 0; (b) a consistent weighting ensures that the SOIL importance of any true variable tends to one as the sample size n goes to infinity; while each variable outside the true model will have the SOIL importance tend to 0.

Theorem 1.

(a) Under the assumption that the weighting $\mathbf{w}$ is weakly consistent, we have:

$$\frac{\sum_{j \in \mathcal{A}^*} S_j}{r^*} \xrightarrow{p} 1, \qquad \frac{\sum_{j \notin \mathcal{A}^*} S_j}{r^*} \xrightarrow{p} 0, \qquad \text{as } n \to \infty;$$

(b) When the weighting $\mathbf{w}$ is consistent, we have:

$$\min_{j \in \mathcal{A}^*} S_j \xrightarrow{p} 1, \qquad \max_{j \notin \mathcal{A}^*} S_j \xrightarrow{p} 0, \qquad \text{as } n \to \infty.$$

In some applications, one may set up a threshold value $c \in (0, 1)$ for the variable importance, and only keeps all the variables whose importances are greater than c. Denote by $\mathcal{A}_c = \{j : S_j > c\}$ the model selected according to this criterion. The property of $\mathcal{A}_c$ is shown in the following theorem, which indicates that for any threshold c, the number of the true variables missed by $\mathcal{A}_c$ and the number of the over-selected variables in $\mathcal{A}_c$ will be relatively small as n grows large.

Theorem 2. For any threshold $c \in (0, 1)$, denote $\overline{\mathcal{A}}_c = \{j \in \mathcal{A}^* : S_j \le c, \quad j = 1, \ldots, p\}$, $\underline{\mathcal{A}}_c = \{j \notin \mathcal{A}^* : S_j > c, j = 1, \ldots, p\}$, then if $\mathbf{w}$ is weakly consistent, we have

$$\frac{|\overline{\mathcal{A}}_c|}{r^*} \xrightarrow{p} 0, \qquad \frac{|\underline{\mathcal{A}}_c|}{r^*} \xrightarrow{p} 0, \qquad \text{as } n \to \infty.$$

As for the choice of threshold, its value depends on how one intends to balance between the cost of overfitting and under-fitting. Actually $|\mathcal{A}_c \nabla \mathcal{A}^*| = |\overline{\mathcal{A}}_c \cup \underline{\mathcal{A}}_c|$. We can also get that $\frac{|\mathcal{A}_c \nabla \mathcal{A}^*|}{r^*} \xrightarrow{p} 0$ as $n \to \infty$. The proofs of Theorem 1 and Theorem 2 are presented in the Appendix.

3. Implementation

3.1. Candidate Models

Now, we discuss how to choose candidate models for computing the SOIL importance. One approach is to use a complete collection of all-subset models as the candidate models, that is,

$$\boldsymbol{A} = \{\varnothing, \{j_1\}, \ldots, \{j_p\}, \{j_1, j_2\}, \{j_1, j_3\}, \ldots, \{j_1, \ldots, j_p\}\},$$

where $j_1, \ldots, j_p \in \{1, \ldots, p\}$. However, in the high-dimensional setting where $p \gg n$, using the candidate models with all subsets is computationally infeasible. Alternatively, we obtain the candidate models using tools for high-dimensional penalized regression

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta}\right)^2 + \sum_{j=1}^{p} p_\lambda(\beta_j), \tag{2}$$

where $p_\lambda(\cdot)$ is a nonnegative penalty function with regularization parameter $\lambda \in (0, \infty)$, such as, Lasso (Tibshirani 1996) penalty $p_\lambda(u) = \lambda w|u|$ in (2), and nonconvex penalties including the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001)

$$p_\lambda(u) = \lambda|u| I(|u| \le \lambda) + \left\{\lambda|u| - \frac{(\lambda - |u|)^2}{2(\gamma - 1)}\right\} I(\lambda < |u| \le \gamma\lambda) + \frac{(\gamma + 1)\lambda^2}{2} I(|u| > \gamma\lambda), \qquad (\gamma > 2),$$

or the minimax concave penalty (MCP, Zhang 2010)

$$p_\lambda(u) = \lambda \left(|u| - \frac{u^2}{2\gamma\lambda}\right) I(|u| \le \gamma\lambda) + \frac{\gamma\lambda^2}{2} I(|u| > \gamma\lambda), \qquad (\gamma > 1).$$

We first apply a high-dimensional model selection method, e.g., SCAD, on the data to compute solution paths for a sequence of tuning parameter $\{\lambda_1, \ldots, \lambda_L\}$. Let $\{\widehat{\boldsymbol{\beta}}^{\lambda_1}, \ldots, \widehat{\boldsymbol{\beta}}^{\lambda_L}\}$ be the estimated coefficients of L different regularization levels for the SCAD penalty and

$$\boldsymbol{A}_{\text{SCAD}} = \{\mathcal{A}^{\lambda_1}, \mathcal{A}^{\lambda_2}, \ldots, \mathcal{A}^{\lambda_L}\}$$

be the resulting models with $\mathcal{A}^{\lambda_l} \equiv \text{supp}(\widehat{\boldsymbol{\beta}}^{\lambda_l}) = \{j : \widehat{\beta}_j^{\lambda_l} \neq 0\}$. We then use the set $\boldsymbol{A}_{\text{SCAD}}$ as the set of candidate models.

To further increase the chance of capturing the true/best model, we can put together the resulting models from several different penalties to form a larger set of candidate models, for example, $\boldsymbol{A} = \{\boldsymbol{A}_{\text{Lasso}}, \boldsymbol{A}_{\text{AdaptiveLasso}}, \boldsymbol{A}_{\text{SCAD}}, \boldsymbol{A}_{\text{MCP}}\}$. The individual penalized methods for producing $\boldsymbol{A}$ do not have to all contain the true model $\mathcal{A}^*$. As long as there is at least one candidate model in the solution paths being (or very close to) the true model, SOIL importance can still work well, provided that the weighing is sensible. By considering multiple model selection methods through merging their solution paths, the chance of including the true model in $\boldsymbol{A}$ is enhanced.

3.2. Weighting

In this article, we focus on two kinds of weighting methods: ARM weighting, which is a weighting strategy by data splitting and cross-assessment, and BIC weighing by BIC or a modified BIC information criterion (BIC-p) for high-dimensional data. Yang and Barron (1998) pointed out that when we have exponentially many models, we may consider the model complexity in terms of the prior weight on the model. When the dimensionality is large, a uniform prior penalty in ARM and BIC does not perform well. Following the same approach in Nan and Yang (2014), we consider a non-uniform prior (or descriptive complexity from a coding perspective) $e^{-\psi C_k}$ when computing both then ARM weighting and the BIC weighting, where ψ is a positive constant and C_k will be given in Algorithm 1.

Weighting using ARM with nonuniform priors. The ARM weighting method randomly splits the data $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of equal size (for simplicity, assume n is an even number). Then, the regression models trained on $\mathbf{D}_1$ are used for prediction on $\mathbf{D}_2$. Then, the weights $\mathbf{w} = (w_1, \ldots, w_K)^\intercal$ can be computed based on this prediction. We consider the linear regression model,

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

Specifically, if we denote by $\boldsymbol{\beta}_s^{(k)}$ the nonzero-coefficient subvector of $\boldsymbol{\beta}^{(k)}$ specified by the model $\mathcal{A}^k$, and let $\mathbf{x}_s^{(k)} \in \mathbb{R}^{|\mathcal{A}^k|}$ be the corresponding subset of predictors, we summarize the ARM weighting method in Algorithm 1.

Algorithm 1 The procedure of the ARM weighting for the regression case.

- Randomly split $\mathbf{D}$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of equal size.
- For each $\mathcal{A}^k \in \boldsymbol{A}$, fit a standard linear regression of y on $\mathbf{x}_s^{(k)}$ using the training set $\mathbf{D}_1$ and get the estimated coefficient $\widehat{\boldsymbol{\beta}}_s^{(k)}$ and the estimated standard deviation $\widehat{\boldsymbol{\sigma}}_s^{(k)}$.
- For each $\mathcal{A}^k$, compute the prediction $\mathbf{x}_s^{(k)\intercal} \widehat{\boldsymbol{\beta}}_s^{(k)}$ on the test set $\mathbf{D}_2$.
- Compute the weight w_k for each candidate model:

$$w_k = \frac{e^{-\psi C_k} \left(\widehat{\boldsymbol{\sigma}}_s^{(k)}\right)^{-n/2} \prod_{i \in \mathbf{D}_2} \exp\left(-\left(\widehat{\boldsymbol{\sigma}}_s^{(k)}\right)^{-2} \left(y_i - \mathbf{x}_{s,i}^{(k)\intercal} \widehat{\boldsymbol{\beta}}_s^{(k)}\right)^2 / 2\right)}{\sum_{l=1}^{K} e^{-\psi C_l} \left(\widehat{\boldsymbol{\sigma}}_s^{(l)}\right)^{-n/2} \prod_{i \in \mathbf{D}_2} \exp\left(-\left(\widehat{\boldsymbol{\sigma}}_s^{(l)}\right)^{-2} \left(y_i - \mathbf{x}_{s,i}^{(l)\intercal} \widehat{\boldsymbol{\beta}}_s^{(k)}\right)^2 / 2\right)},$$

  for $k = 1, \ldots, K$, where $C_k = s_k \log \frac{e \cdot p}{s_k} + 2\log(s_k + 2)$ and $s_k = |\mathcal{A}^k|$ is the number of non-constant predictors for model k.
- Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$ for $l = 1, \ldots, L$, and get $w_k = \frac{1}{L} \sum_{l=1}^{L} w_k^{(l)}$.

Weighting using information criteria with nonuniform priors. An alternative way of weighting is using BIC information criteria. Define $I_k^{\text{BIC}} = -2 \log \ell_k + s_k \log n$ as the BIC information

criterion, where ℓ_k is the maximized likelihood for model k and $s_k = |\mathcal{A}^k|$ denotes the number of nonconstant predictors. Then weight w_k for model $\mathcal{A}^k \in \boldsymbol{A}$ is computed by

$$w_k = \exp\left(-\frac{I_k}{2} - \psi C_k\right) \Big/ \sum_{l=1}^{K} \exp\left(-\frac{I_l}{2} - \psi C_l\right). \quad (3)$$

We refer to the above approach with nonuniform priors as the BIC-p weighting.

Besides the ARM and BIC-p weighting, one can also consider another alternative weighting approach by using Fisher's fiducial idea from the generalized fiducial inference (Lai, Hannig, and Lee 2015). The details are included in supplementary materials Part A. We do not discuss this method in details since it only applies to the regression settings.

Often consistency of a weighting method is proved when all subset models are considered (e.g., Lai, Hannig, and Lee 2015). But when p is large, it is computationally infeasible to include all the variables, so some screening methods may be applied to reduce the number of variables. Next, we prove that under certain assumptions, SOIL importance is consistent on differentiating important variables from unimportant ones:

Corollary 1. Under the assumption that the weighting $\mathbf{w}$ on the all-subset candidate models $\boldsymbol{A}$ is consistent, as long as at least one method is path-consistent, we have

$$\min_{j \in \mathcal{A}^*} S(j; \mathbf{w}', \boldsymbol{A}') \xrightarrow{p} 1, \quad \max_{j \notin \mathcal{A}^*} S(j; \mathbf{w}', \boldsymbol{A}') \xrightarrow{p} 0, \quad \text{as } n \to \infty,$$

where $\mathbf{w}'$ is the renormalized weighting on $\boldsymbol{A}'$, which is the collection of models using union of solution paths.

3.3. Software

We provide our implementation of the SOIL importance measure in an official R package `SOIL`, which is publicly available from the Comprehensive R Archive Network at *https://cran.r-project.org/web/packages/SOIL/index.html*. The package is also provided in the supplementary materials.

4. Extension to the Binary Classification Model

We extend the SOIL importance to the binary logistic regression case. Let $Y \in \{0, 1\}$ be the response variable and $X \in \mathbb{R}^p$ be the predictor vector. We assume that Y has a Bernoulli distribution with conditional probabilities

$$\Pr(Y = 1|X = \mathbf{x}) = 1 - \Pr(Y = 0|X = \mathbf{x}) = \frac{e^{\mathbf{x}^\intercal \boldsymbol{\beta}^*}}{1 + e^{\mathbf{x}^\intercal \boldsymbol{\beta}^*}}, \quad (4)$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\intercal$ is the vector corresponding to the true underlying model. The ARM weighting for the logistic regression can be computed by Algorithm 2.

Algorithm 2 The procedure of the ARM weighting for the binary classification case.

- Randomly split $\mathbf{D}$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of equal size.
- For each $\mathcal{A}^k \in \boldsymbol{A}$, fit a standard logistic regression of y on $\mathbf{x}_s^{(k)}$ using the samples in $\mathbf{D}_1$. Obtain the estimated coefficients $\widehat{\boldsymbol{\beta}}_s^{(k)}$ and the corresponding function of predicted conditional probability:

$$\widehat{p}^{(k)}(\mathbf{x}) \equiv \Pr(Y = 1|X_s^{(k)} = \mathbf{x}) = \exp\left(\mathbf{x}^\intercal \widehat{\boldsymbol{\beta}}_s^{(k)}\right) / \left(1 + \exp\left(\mathbf{x}^\intercal \widehat{\boldsymbol{\beta}}_s^{(k)}\right)\right), k = 1K.$$

- For each $\mathcal{A}^k$, compute the predicted probability $\widehat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})$ on the test set $\{i | i \in \mathrm{D}_2\}$.
- Compute the weight w_k for each candidate model:

$$w_k = \frac{e^{-\psi C_k} \prod_{i \in \mathbf{D}_2} \widehat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})^{y_i} \left(1 - \widehat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})\right)^{1-y_i}}{\sum_{l=1}^{K} e^{-\psi C_l} \prod_{i \in \mathbf{D}_2} \widehat{p}^{(l)}(\mathbf{x}_{s,i}^{(l)})^{y_i} \left(1 - \widehat{p}^{(l)}(\mathbf{x}_{s,i}^{(l)})\right)^{1-y_i}},$$

  for $k = 1, \ldots, K$, where $C_k = s_k \log \frac{e \cdot p}{s_k} + 2\log(s_k + 2)$ and $s_k = |\mathcal{A}^k|$ is the number of nonconstant predictors for model k.
- Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$ for $l = 1, \ldots, L$, and get $w_k = \frac{1}{L} \sum_{l=1}^{L} w_k^{(l)}$.

4.1. Weighting Using Information Criteria with Nonuniform Priors

Similarly, the weight w_k for model $\mathcal{A}^k \in \boldsymbol{A}$ using BIC-p the information criterion can be computed in the same way as in (3) where $I_k^{\mathrm{BIC}} = -2\log \ell_k + 2s_k \log n$, with $s_k = |\mathcal{A}^k|$ and ℓ_k being the maximized likelihood function for the logistic model $\mathcal{A}^k$.

5. Simulations

In this section, we consider a number of simulation settings to highlight the properties of SOIL in contrast to some other importance measures. We compare SOIL using the ARM and BIC-p weighting methods with three variable importance alternatives, which are denoted as LMG, RFI1, and RFI2. LMG is the relative importance measure by averaging over all possible orderings for R^2 decomposition (Lindeman, Merenda, and Gold 1980). RFI1 and RFI2 are importance measures in random forests proposed by Breiman (2001). Specifically, RFI1 is computed from a normalized difference between the prediction error on the out-of-bag (OOB) portion of the data and that on the permuted OOB data for each predictor variable. RFI2 is the total decrease in node impurities from splitting on a particular variable, averaged over all trees. The node impurity is defined by the Gini index for classification, and by residual sum of squares for regression. Computationally, LMG can be obtained by the R implementation `relaimpo` (Grömping et al. 2006), while RFI1 and RFI2 can be obtained by R implementation `randomForest` (Liaw and Wiener 2002). Since LMG can only handle the linear case with up to about 20 variables due to its computational limitation, we are not able to get the relative importance LMG in some of our examples. In all the simulations, we obtain $\boldsymbol{A}_{\text{lasso}}$, $\boldsymbol{A}_{\text{SCAD}}$ and $\boldsymbol{A}_{\text{MCP}}$ separately on the whole dataset under the default settings of the tuning parameters from the package `glmnet` (lasso) and `ncvreg` (SCAD and MCP), respectively. Then, we use the union of $\boldsymbol{A}_{\text{lasso}}$, $\boldsymbol{A}_{\text{SCAD}}$, and $\boldsymbol{A}_{\text{MCP}}$ as our candidate set $\boldsymbol{A}$.

Table 1. Simulation settings.

Example	n	p	Model settings
			Gaussian case
1	100	1000	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$
2	150	14+1	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, \ldots, 0)^\intercal$. Add $X_{15} = 0.5X_1 + 2X_4 + e$ and $\beta^*_{15} = 0$, where $e \sim N(0, 0.01)$.
3	150	8	$\boldsymbol{\beta}^* = (0, \ldots, 0)^\intercal$
4	150	8	$\boldsymbol{\beta}^* = (1, \ldots, 1)^\intercal$
S1	150	20	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, \ldots, 0)^\intercal$
S2	150	6+6	$\boldsymbol{\beta}^* = (4, 4, -6\sqrt{2}, \frac{4}{3}, 0, 0)^\intercal$. Add $(X_1^2, X_2^2, X_3^2, X_4^2, X_5^2, X_6^2)$ and corresponding coefficients $(\beta^*_7, \beta^*_8, \ldots, \beta^*_{12})^\intercal = (4, 0, 1, 0, 0, 0)^\intercal$.
S3	150	6+6	$\boldsymbol{\beta}^* = (4, 4, -6\sqrt{2}, \frac{4}{3}, 0, 0)^\intercal$. Add $(X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4)$ and corresponding coefficients $(\beta^*_7, \beta^*_8, \ldots, \beta^*_{12})^\intercal = (4, 2, 2, 0, 0, 0)^\intercal$.
S6	100	200	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$
			Binomial case
5	80	6	$\boldsymbol{\beta}^* = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0)^\intercal$
6	5000	6	$\boldsymbol{\beta}^* = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0)^\intercal$
S4	150	20	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$
S5	100	200	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$

In the following, we compare different variable importance measures for Gaussian and Binomial cases under various settings of sample sizes, dimensions, and feature correlations.

Model 1: Gaussian. The simulation data $\{y_i, \mathbf{x}_i\}_{i=1}^n$ are generated from the linear model $y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma \in \{0.1, 5\}$. We generate $\mathbf{x}_i$ from multivariate normal distribution $N_p(0, \Sigma)$. For each element Σ_{ij} of Σ, $\Sigma_{ij} = \rho^{|i-j|}$, that is, the correlation of X_i and X_j is $\rho^{|i-j|}$, with $\rho \in \{0, 0.9\}$.

Model 2: Binomial. The iid sample $\{y_i, \mathbf{x}_i\}_{i=1}^n$ is generated from the binomial model $\text{logit}(p_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta}^*$, where $p_i = P(Y = 1|X = \mathbf{x}_i)$. And $\mathbf{x}_i$ is generated in the same way as the Gaussian case.

We summarize in Table 1 the model settings adopted in this simulation. For each model setting with a specific choice of the parameters (ρ, σ^2), we repeat the simulation 100 times and compute the averaged variable importance measures for SOIL-BIC-p, SOIL-ARM, LMG, RFI1, and RFI2.

The results for the simulations are shown in Figures 1–6 and Figures S1–S6. Due to page restrictions, the figures of Example S1–S6 are only provided in the supplementary materials, while the summary of all the examples are discussed in the main part of the article. For the scaling of the importance measures, we standardize RFI1 and RFI2, dividing them by their respective maximum value of the variable importance among all the

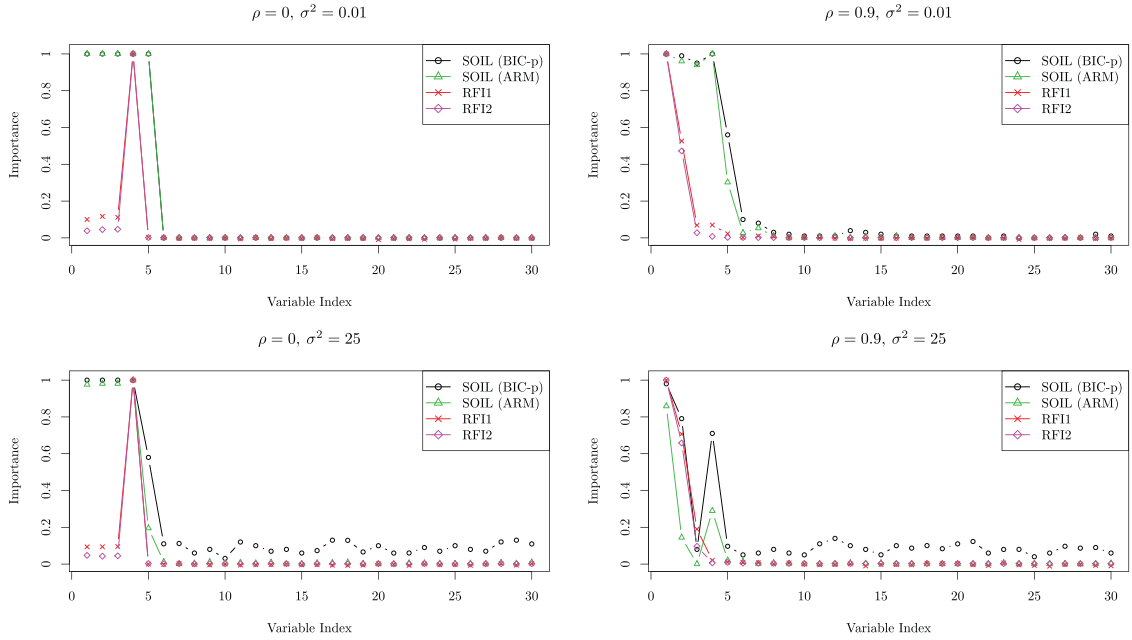

Figure 1. Simulation results for Example 1, where $n = 100$, $p = 1000$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, \ldots, 0)$.

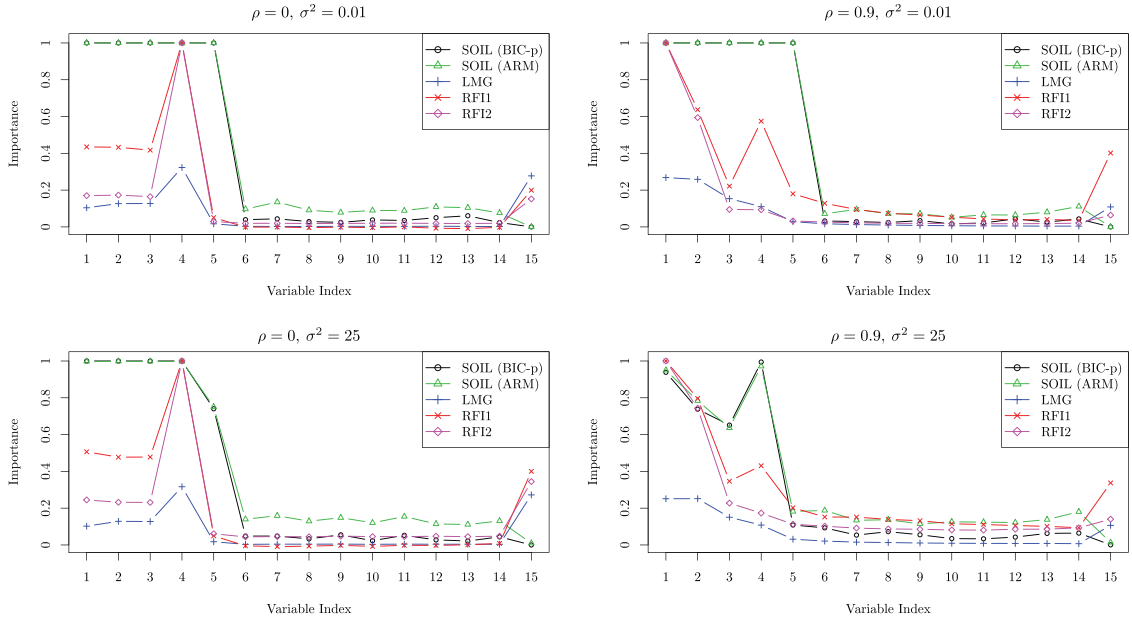

Figure 2. Simulation results for Example 2, where $n = 150$, $p = 14$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, \ldots, 0)$. Add $X_{15} = 0.5 * X_1 + 2 * X_4 + e$ and corresponding $\beta^*_{15} = 0$, where $e \sim N(0, \sigma_e^2)$.

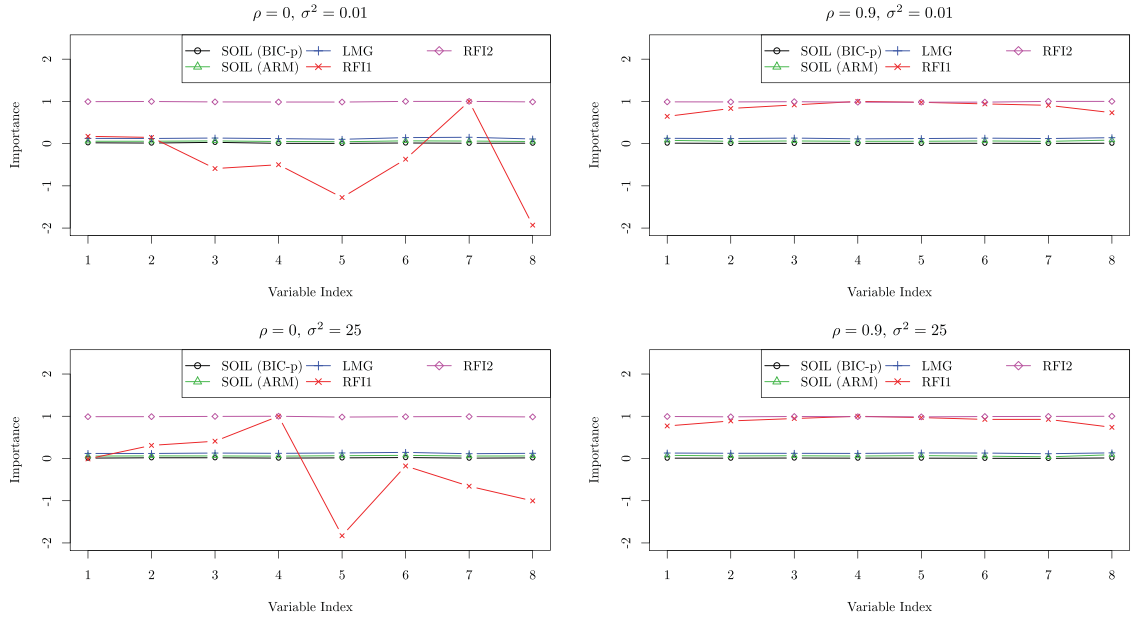

Figure 3. Simulation results for Example 3, where $n = 150$, $p = 8$. The true coefficients $\boldsymbol{\beta}^* = (0, \ldots, 0)^\top$.

variables for each realization of the data. As a result, in each figure, we can see that the maximum value of RFI1 or RFI2 (after the standardization) is always one. For SOIL and LMG, we keep their original values as being proposed. The fact that the LMG importance values sum to one over the variables should be kept in mind when comparing the different importance measures on the graphs.

The choice of the prior ψ for the ARM and BIC-p weighting can be specified by the users. To avoid cherry-picking, we present the results with a fixed choice: $\psi = 0.5$. Our experience is that $\psi = 0.5$ or 1 generally works quite well. We conduct a sensitivity analysis on the choice of ψ, which is presented in Figure S6 in the Supplementary Materials. We tried eight different values, that is, $\psi \in \{0, 0.5, 1, 1.5, 2, 3, 3.5, 10\}$ on the low noise ($\sigma^2 = 0.01$) and high correlation ($\rho = 0.9$) case of Example S6. We can conclude that a too large value $\psi = 10$ leads to poor performance of SOIL, that is, detecting nothing important, while choices of too small ψ (0 or close to 0) may result in significant SOIL importances of unimportant variables. Overall, SOIL importances under $\psi = 0.5$ or $\psi = 1$ are stably reliable in our simulations.

5.1. Relative Performances of Importance Measures in Several key Aspects

A summary of the relevant properties of different important measures is provided in Table 2. In the following, we discuss point-by-point these characteristics for the importance measures in comparison. For convenience, we call the variables with nonzero coefficients the "true" variables.

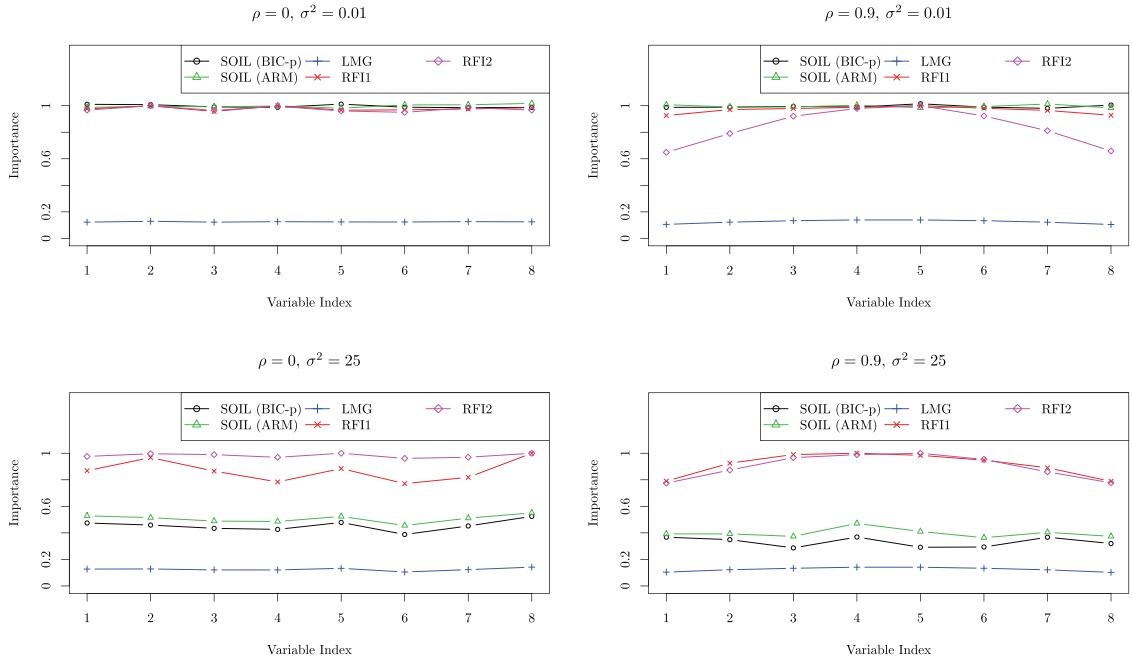

Figure 4. Simulation results for Example 4, where $n = 150$, $p = 8$. The true coefficients $\boldsymbol{\beta}^* = (1, \ldots, 1)^\intercal$.

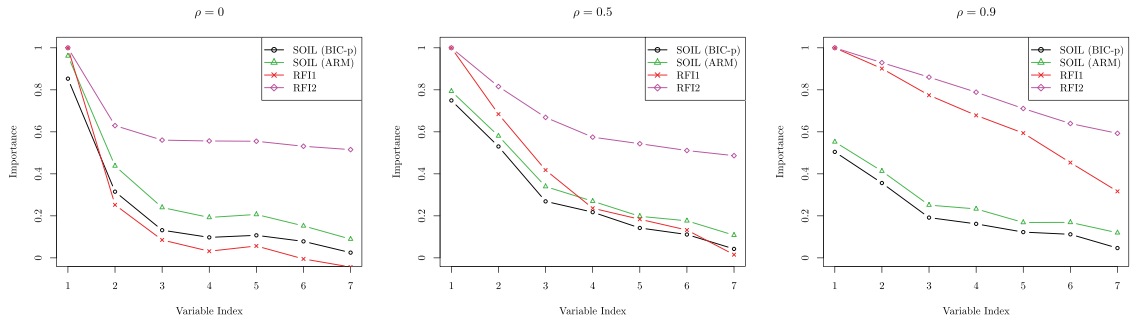

Figure 5. Simulation results for Example 5, where $n = 80$, $p = 6$. The true coefficients $\boldsymbol{\beta}^* = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0)^\intercal$.

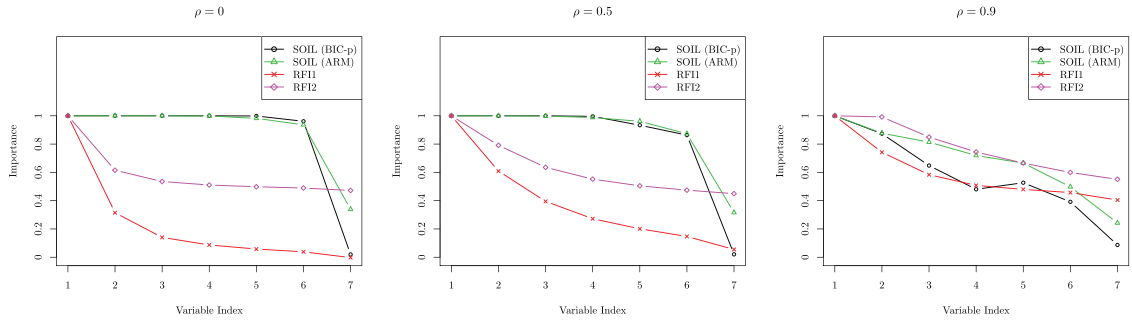

Figure 6. Simulation results for Example 6, where $n = 5000$, $p = 6$. The true coefficients $\boldsymbol{\beta}^* = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0)^\intercal$.

Inclusion/exclusion. The inclusion/exclusion aspect addresses the issue if an importance measure can give a proper sense if a predictor is likely to be needed in the best model to describe the data. These two criteria for importance have been discussed in Grömping (2015). Recall that given enough data for SOIL importance, the true variables in the model have large importances (inclusion) and the variables that are not in the true model have importances around zero (exclusion). In all examples, we can see that the SOIL-BIC-p and SOIL-ARM have the inclusion/exclusion property. For example in Figure S1, all the true variables $(X_1, \ldots, X_5)$ have their SOIL importances around one, even though their coefficients are different, i.e. $(\beta_1^*, \ldots, \beta_5^*) = (4, 4, 4, -6\sqrt{2}, \frac{4}{3})$. In contrast, the other three measures LMG, RFI1, and RFI2 do not have the inclusion property when $\rho = 0$ and $\sigma^2 = 0.01$ (they all undervalue the importance of X_5, which has a small coefficient). LMG, RFI1,

and RFI2 do not have the exclusion property either. We can see that in Figure 2 the noise variable X_{15} confuses LMG, RFI1, and RFI2. In Figure S2 when $\rho = 0.9$, LMG, RFI1, and RFI2 assign relatively high values on the noise variable X_8. In Figure S3, when $\rho = 0.9$ and $\sigma^2 = 25$, LMG, RFI1, and RFI2 fail on the noise variable X_{10}.

SOIL is certainly incapable of giving high importance to very weak variables in the true model. For example Figure 5 shows that in a binomial model with the decreasing coefficient vector $\boldsymbol{\beta}^* = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0)^\intercal$, the true variable X_6's SOIL importance is only around 0.1, not much above that of the noise variable X_7). However, this problem is alleviated as the sample increases: Figure 6 shows that the SOIL-ARM and SOIL-BIC-p importances of six true variables $(X_1, \ldots, X_6)$ become closer to one when n increases from 80 to 5000. In contrast, the LMG, RFI1, and RFI2 stay basically the same as the sample size increases.

Tuning in to information. For high-dimensional data, more often than not (to say the least), sparsity is a reluctant acceptance that the info and/or computational limit only allows us a simple model for application. The optimal sparsity should depend on the sample size and noise level. Therefore, it is desirable to have an importance measure to honor this perspective. When the sample size increases or the noise decreases, we should have more information. Thus, the importance obtained from the data should change due to the enrichment of information. Therefore in most examples, when the correlation ρ and σ^2 are low, one may hope the variable importances delineate the true model. Comparing Examples 5 and 6, which differ only in the sample size, as shown in Figure 5 and 6, only SOIL-BIC-p and SOIL-ARM react to the much increased information due to sample size increase, while the other three importances are not tuned in to the information change.

Robustness to feature correlation. SOIL importances show robustness against noise increase and higher feature correlation. For example in Figure 1, 2, and Figures S1–S5 in Supplementary Materials Part B, even when there is high feature correlation ($\rho = 0.9, \sigma^2 = 0.01$) or strong noise ($\rho = 0, \sigma^2 = 25$) in the data, the SOIL-BIC-p and SOIL-ARM still give relatively large importance values to the true variable X_5, while the other methods consider X_5 as unimportant. But in a case of both high feature correlation and strong noise ($\rho = 0.9, \sigma^2 = 25$), none of the importance measures in comparison can quite clearly select X_5 as an important variable because the information is too limited.

Table 2. Comparison of the characteristics for the importance measures. A "√" indicates that a specified method has the given property. A blank space indicates the absence of a property.

	SOIL-ARM	SOIL-BIC-p	LMG	RFI1	RFI2
Inclusion/Exclusion	√	√			
Tuning in to information	√	√			
Robustness to feature correlation	√	√			
Robustness against confuser	√	√			
Sensitivity to high-order terms	√	√			
Pure relativeness			√	√	√
Order preserving	√	√			
High-dimensionality	√	√		√	√
Non-parametricness				√	√
Non-negativity	√	√	√		√

Robustness against confusers. A confuser refers to a variable that is closely related to a true variable or some linear combination of the true variables but not to the extent of serving as a valid alternative. An importance measure oriented towards sparse modeling should assign near zero importances on the confusers. The simulation results show that the SOIL importance measures are much more robust to confusers than LMG, RFI1, and RFI2. In Example 2, we generate a confuser $X_{15} = 0.5X_1 + 2X_4 + e$ with Gaussian noise $e \sim N(0, 0.01)$. The results in Figure 2 show that LMG, RFI1 and RFI2 fail to assign small importance to X_{15} (not in the true model) and view it more important than some true variables. In contrast, small ARM and BIC-p importances for X_{15} correctly indicate that it is unimportant.

Sensitivity to higher-order terms. The SOIL importance measures are more sensitive to inclusion of higher-order terms in the model. In Example S2 and S3, we add quadratic terms X_1^2, $X_2^2, X_3^2, X_4^2, X_5^2, X_6^2$ and pairwise interactions X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4 respectively, where the coefficients for X_1X_2, X_1X_3, X_1X_4 and X_1^2, X_3^2 are nonzero in the true models. Results in Figure S2 and S3 show that the ARM and BIC-p methods can select both true main-effect variables and true higher-order terms, whereas LMG, RFI1, and RFI2 fail to select some of the main-effect variables when interactions or quadratic terms are included.

Pure relativity. An importance measure is said to be purely relative if the values individually do not have a sensible meaning on their own. One drawback of an importance measure with pure relativity is that it does not differentiate between equal importance and equal unimportance cases. All coefficients in Example 3 and 4 have the same relative size, which are $\boldsymbol{\beta}^* = (0, \ldots, 0)^\intercal$ and $\boldsymbol{\beta}^* = (1, \ldots, 1)^\intercal$, respectively. We find that LMG, RFI1, and RFI2 do not offer any clue on importance of each variable itself. Variables $(X_1, \ldots, X_6)$ in Example 3 have very similar LMG and RFI2 values to those in Example 4. And RFI1 behaves wildly as it assigns very much different importances to the variables in the independence case ($\rho = 0$) of Example 3. The importance values are even significantly negative for some variables. In contrast, SOIL-BIC-p and SOIL-ARM nicely separate the two examples.

Order preserving. Order preserving refers to the property that the importance reflects the "order" of the variables or not: (1) for the true variables (standardized) with not too high correlations with others, it may be natural to expect the ones with larger coefficients to have larger importances (up to one of course); (2) the true variables should have larger importances compared to the noise ones. In the case that the sample size is too small for some true variables to be detectable, the order preserving property demands that the noise variables should not receive significantly higher importance values than these subtle true variables. SOIL-BIC-p and SOIL-ARM exhibit the order preserving property in all the cases. LMG behaves poorly when there exists a confuser as in Figure 2. RFI1 and RFI2 do not preserve the order when correlation $\rho = 0.9$ and/or noise σ^2 is large.

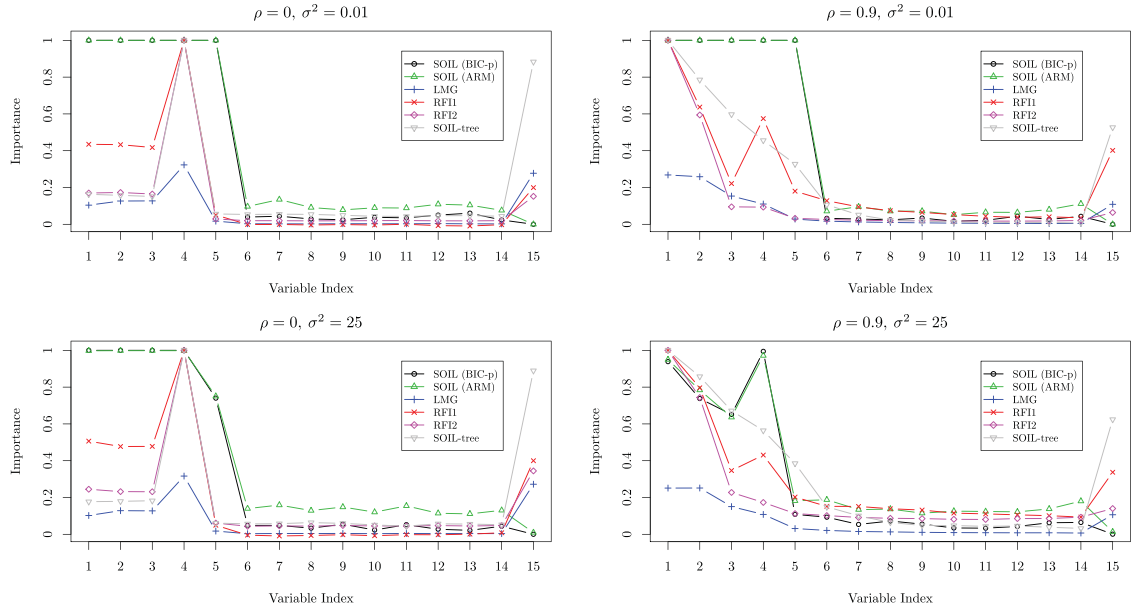

Figure 7. Simulation results for SOIL-tree on Example 2.

High-dimensionality. SOIL-BIC-p, SOIL-ARM, RFI1, and RFI2 can work for high-dimensional data when $p > n$ as shown in Figure 1 and S5. The exclusion and inclusion properties still hold for SOIL-BIC-p and SOIL-ARM in the high-dimensional case (inclusion of a weak variable requires that σ^2 is not too high). In contrast, LMG does not support high-dimensional data.

Nonnegativity. SOIL-BIC-p, SOIL-ARM, LMG, and IMG2 always yield nonnegative importance value. However, RFI1 does not satisfy this criterion.

Nonparametricness. Among the importance measures, only the two from random forest are not limited to parametric modeling.

5.2. Comparison of SOIL with Lasso and Stability Selection

Meinshausen and Bühlmann (2010) proposed a stability selection (SS) method to improve the Lasso variable selection. SS may be regarded as an importance measure. In Supplementary Materials Part C, we present a comparison of SS importance to our SOIL approach. Additionally, in Supplementary Materials Part D, we present a stability comparison of Lasso and SOIL. Due to the worse performances of SS and Lasso compared with SOIL, together with the fact that the main goals of SS and Lasso are not on variable importance, we do not consider SS or Lasso in our main simulation.

5.3. Influence of the Weighting Method on tree Models

Are the advantages of the SOIL approach compared to random forest seen so far mainly due to the data driven model averaging instead of the simple averaging as in random forest? We here investigate the SOIL type weighting on the tree models. Like the BIC weighting methods, we use the cost complexity of a tree, $I_\alpha(T_k) = \Sigma_{m=1}^{|T|} N_m Q_m(T_k) + \alpha|T_k|$, to calculate the weights for the kth tree T_k, where $|T_k|$ is the number of terminal nodes in the tree T_k, N_m is the number of observations in each terminal of the tree, α is the tuning parameter (selected by cross-validation) and $Q_m(T_k)$ is the deviance (node impurity if it is a classification tree) of the mth terminal node in T_k. Every tree produces a list of variable importance and we use the weighted sum of these lists of tree variable importances as the final importance measure, which we call SOIL-tree. We apply this measure in Example 2. Figure 7 shows the results. Comparing the SOIL-ARM/BIC-p with SOIL-tree, we can see the SOIL-ARM/BIC-p perform better than SOIL-tree in differentiating the true important variables. Comparing the RFI1/RFI2 with SOIL-tree, we see that the SOIL weighting improves the performances of random forest in the high correlation high noise case. The former comparison indicates that the differences between SOIL and RF1/RF2 goes beyond the weighting difference in SOIL and random forest and the latter suggests that the SOIL weighting strategy can improve the performance of tree-model based importances in the high-correlation and high-noise case.

6. Real Data Examples

We apply the variable importance measures to three real datasets:

BGS data. We first consider a dataset with small p from the Berkeley Guidance Study (BGS) by Tuddenham and Snyder (1954). The dataset includes 66 registered newborn boys whose physical growth measures are followed for 18 years. Following Cook and Weisberg (2009, p. 179) we consider a regression model of age 18 height on $p = 6$ predictors: weights at ages two (WT2) and nine (WT9), heights at ages two (HT2) and nine (HT9), age nine leg circumference (LG9), and age 18 strength (ST18). The corresponding SOIL-ARM, SOIL-BIC-p, LMG, RFI1, and RFI2 importances for each variable are computed and summarized in Table 3. We found that HT9 is the most important variable according to all methods. But different methods produce different second-most important variables.

Then, we conduct a "credibility check" for the above results of various importance measures. To do so, we use a guided simulation or cross-examination (Li, Lue, and Chen 2000; Rolling

Table 3. Importance measures of the variables in BGS data. The top two most important variables according to each measure are in bold.

	WT2	HT2	WT9	HT9	LG9	ST18
SOIL-ARM	0.16	0.09	0.03	**1.00**	**0.62**	0.28
SOIL-BIC-p	0.01	0.00	0.00	**1.00**	**0.63**	0.08
LMG	0.06	**0.13**	0.08	**0.65**	0.05	0.02
RFI1	1.72	2.50	1.79	**55.66**	**4.12**	1.05
RFI2	70.89	101.58	100.52	**2126.64**	123.52	**127.74**

and Yang 2014), in which the performances of the importance measures are tested using data that are simulated from models recommended by the importance measures respectively. The basic idea of cross-examination is that one usually anticipates that a good method should have a better performance than other methods on the simulated data that are constructed from the method itself. In our context, if we compute the variable importances $S_1^A, \ldots, S_p^A$ on a real dataset using measure A, and construct a suggested model (with top rated important variables) and simulate a new dataset from this model, then on the new dataset, the variable importances $\tilde{S}_1^A, \ldots, \tilde{S}_p^A$ using measure A should be more similar to $S_1^A, \ldots, S_p^A$ than the variable importances $\tilde{S}_1^B, \ldots, \tilde{S}_p^B$ using measure B. Otherwise, one can naturally question the adequacy of applying measure A to the original real data.

The cross-examination procedure is as follows:

1. Choose one measure from SOIL-ARM, SOIL-BIC-p, LMG, RFI1, and RFI2 as the base measure, and select the resulting top two most important variables (e.g., HT9 and LG9 if SOIL-ARM is the base measure).
2. Fit linear regression using only the selected variables as predictors, and obtain the estimated coefficients $\widehat{\boldsymbol{\beta}}$ and standard deviation $\widehat{\sigma}$.
3. Generate the new response according to the model: $\mathbf{Y}_{\text{new}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\sigma}N(0, 1)$.
4. Compute the SOIL-ARM, SOIL-BIC-p, LMG, RFI1, and RFI2 importance measures using the new dataset $(\mathbf{X}, \mathbf{Y}_{\text{new}})$.
5. Repeat the above steps 100 times and take the average of each importance.
6. Go to Step 1 until all measures have served as the base measure.

Table 4. Classical significance (p-value) analysis of the BGS data.

	Intercept	WT2	HT2	WT9	HT9	LG9	ST18
p-Value	2E−16	0.112	0.105	0.773	4.93E−16	0.246	0.258

The results are depicted in Figure 8. Overall, SOIL-ARM and SOIL-BIC-p perform reasonably better than the other importance measures. In the home-game (where the variables are selected based on the base measure) of SOIL-ARM, SOIL-BIC-p, and RFI1, we can see that LMG and random forest (RFI1 or RFI2) do not support the true variable LG9, while SOIL-ARM or SOIL-BIC-p clearly indicate, correctly, HT9 and LG9 as the important ones (although with less confidence on LG9). In fact, LMG, RFI1, and RFI2 all view HT2 as more important than LG9, a mistake seemingly caused by the higher correlation of HT2 (0.57) to HT18 than LG9 (0.37). In the home-game of LMG, all methods single out only HT9 as the most important (but not HT2). However, SOIL-ARM and SOIL-BIC-p assign the second largest importance to HT2, which is consistent with the aforementioned Order Preserving property. The random forest importance measures do not show this property. The home-game of RFI2 is similar to the home-game of LMG, where the Order Preserving property still holds for SOIL-ARM and SOIL-BIC-p but not for the others.

We also perform a linear regression analysis on the full model directly in the BGS application. The p-values for the variable are presented in Table 4. If we compare the p-values with significance level $\alpha = 0.1$, the only significant variables are the intercept and "HT9". Consistently, HT9 is declared important according to all the variable importances we considered. In terms of p-value, HT2 is the second most important variable, which agrees with LMG, but is different from both the random forest and SOIL importances in Table 3. Based on the earlier guided simulation results, together with the intuition that given HT9, HT2 is unlikely to be that useful for predicting height at age 18, we tend to think the significance analysis based on the full model is less trustworthy. In general, as is well-known, p-value can be quite sensitive to the model used to fit the data, and thus may not be reliable to measure variable importance.

Bardet Data. For a dataset with large p, we consider the Bardet dataset. It collects tissue samples from the eyes of 120

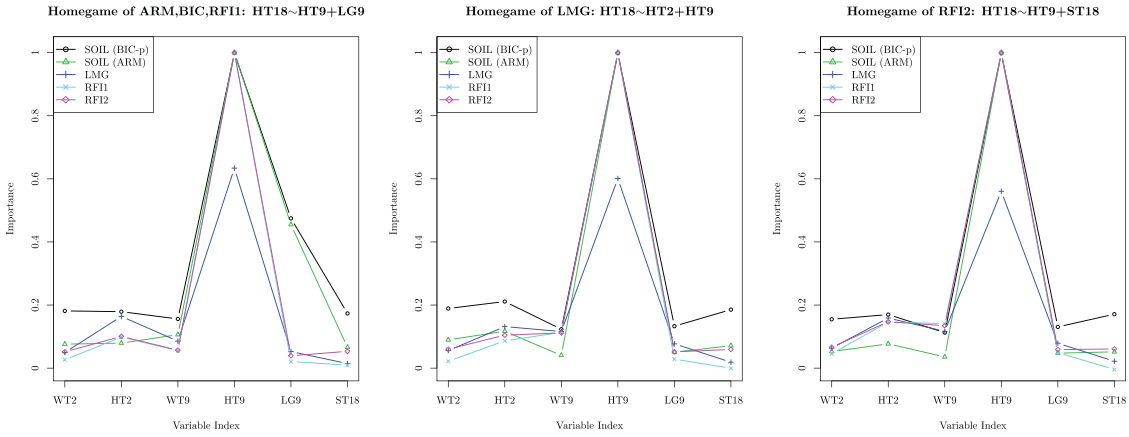

Figure 8. Results of cross-examination for BGS data.

Table 5. Top ten genes for different variable importance measures for Bardet data.

Rank	ARM		BIC-p		RFI1		RFI2	
1	**25141**	1.000	**25141**	1.000	**25141**	5.113	**21907**	0.061
2	**28967**	0.935	**28967**	1.000	**21907**	5.006	**25141**	0.059
3	**28680**	0.834	**28680**	0.999	**11711**	4.875	**11711**	0.054
4	**30141**	0.576	**30141**	0.491	**11719**	4.778	**25105**	0.041
5	21092	0.397	21092	0.278	**25105**	4.491	**24565**	0.036
6	15863	0.261	15863	0.142	**9303**	4.332	**28680**	0.035
7	17599	0.219	17599	0.121	**28680**	4.239	**25403**	0.034
8	22813	0.106	25367	0.028	**25425**	3.788	**9303**	0.033
9	25367	0.079	22813	0.016	**16569**	3.733	**22029**	0.032
10	24892	0.047	14949	0.005	**22029**	3.680	**24087**	0.030

twelve-week-old male rats, which are the offspring of intercrossed F1 animals. For each tissue, the RNAs of 31,042 selected probes are measured by the normalized intensity valued. The gene intensity values are in log scale.

To investigate the genes that are related to gene TRIM32, which causes the Bardet–Biedl syndrome according to Chiang et al. (2006), a screening method (Huang, Ma, and Zhang 2008) is applied to the original probes, which gives us a dataset with 200 probes for each of 120 tissues. Specifically, 3000 out of the 31,042 probes are selected with the largest variances. Then, we select 200 probes with the largest marginal correlation with the response TRIM32 to obtain the reduced dataset, which is available upon request. We use this screened dataset to carry out our importance measure analysis.

Since LMG is not feasible to handle cases with $p > 20$, it is not included in our analysis below. The corresponding SOIL-ARM, SOIL-BIC-p, RFI1, and RFI2 importances for most important variable are summarized in Table 5. We present the top ten variables according to the different importance measures, respectively. The name of each gene is too long, so for convenience we record the corresponding EST number instead. From Table 5, we can see that different importance measures have very different results.

Notice that X_{25141} is the most important variable according to Table 5. Random forest is unstable in the sense that each time we compute the random forest importance on the data, the top ten variables obtained tended to be quite different in terms of their rankings. For SOIL-BIC-p and SOIL-ARM, the top four genes always have the same rank and the importance values are pretty much the same in different runs. Also, a striking feature for the random forest in this data example is that the values of the importances are quite close to each other and decaying gradually, making it hard to judge which variables are really important.

We carry out a guided simulation study similar to that for the BGS data, except that LMG is not included. Based on the information in Table 5, the top 4 variables are selected for SOIL-BIC-p (SOIL-ARM), and the top 10 for RFI1 and RFI2, respectively.

In Figure 9, we only present the variable importances of the "true" genes due to space limitation. RFI1 and RFI2 are all normalized. In the home-game of SOIL-ARM and SOIL-BIC-p, both can correctly select all the true variables if the cut-off value is set at 0.4. For random forest, however, the maximum RFI1 and RFI2 values among the unimportant ones exceed the most important ones respectively, indicating that the random forest has difficulty differentiating the really important and unimportant variables.

In the home-game of RFI1 and RFI2, none of the competitors performs very well. With the generating model being larger, with the limited information in the data (in conjunction with the complicated correlation among the genes), the importance measures simply cannot reveal all the true variables. Only the true variable X_{25414} is differentiated clearly by all methods. From the SOIL perspective, it is willing to support at most three more variables with some confidence. Random forest gives more true variables significant importance values. A drawback is that some noise variables receive relatively large importance values, which are even higher than almost half of the true variables.

From the guided simulations, the Order Preserving property fails in all the cases for the random forest importance measures. For SOIL, in the home-game of ARM and BIC-p, it holds for

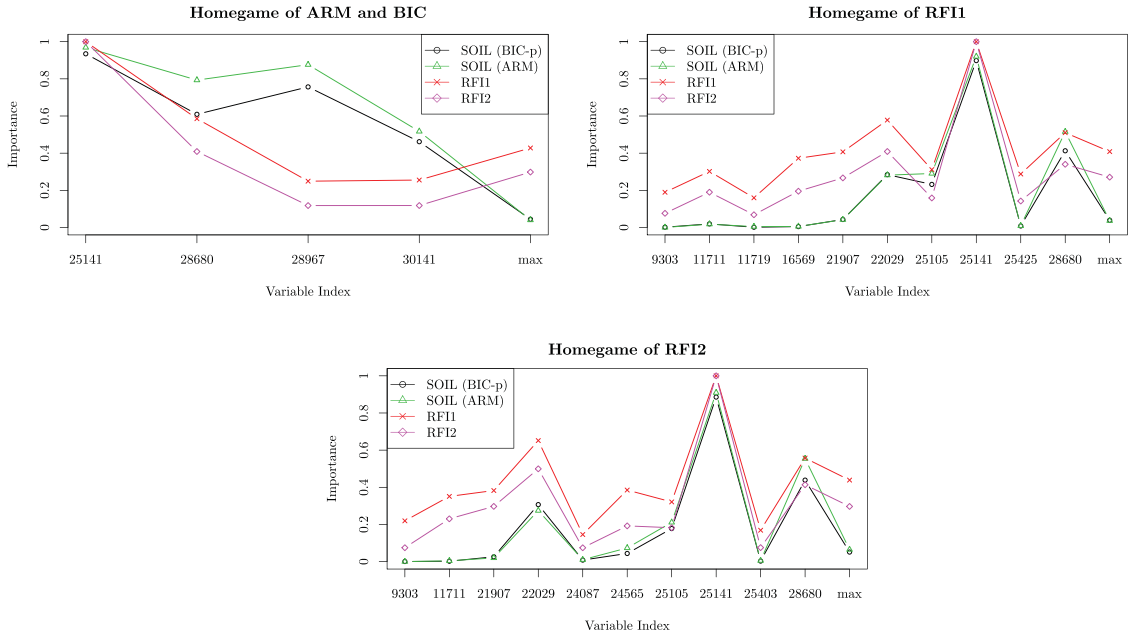

Figure 9. Simulation results for cross-examination.

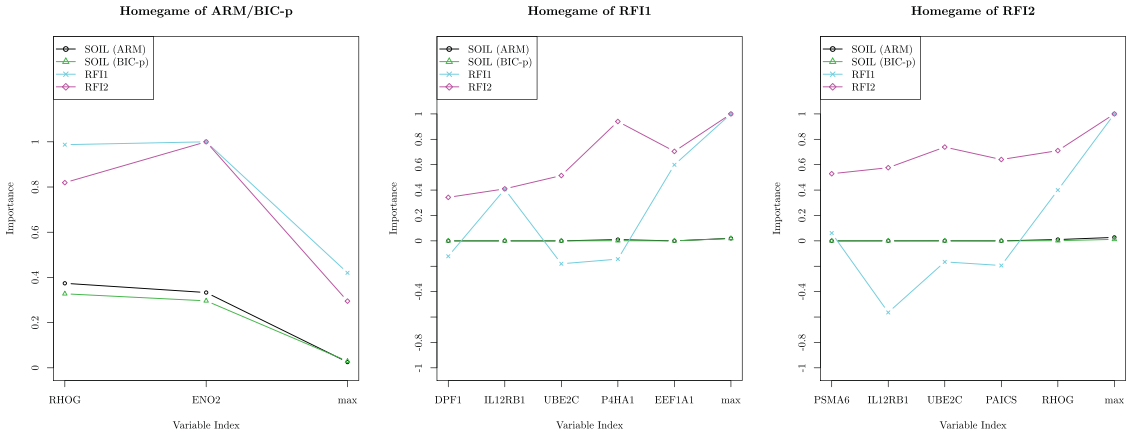

Figure 10. Results of cross-examination for lung cancer data.

both SOIL-ARM and SOIL-BIC-p; but in the home-game of RFI1 or RFI2, the property does not hold exactly, but it does hold in the sense that the maximum importance of the noise variables is still very small (and it is not meaningful to rank the variables with tiny importance values). The key point here is that while SOIL certainly can miss subtle variables in the true model when the sample size is small, it typically does not recommend an unimportant variable as important. The same cannot be said for the other importance measures.

Lung Cancer Data. We analyze a lung cancer gene expression dataset (Subramanian et al. 2005) with 62 patients and 5217 genes. As more and more genomics studies have been done, analyzing and interpreting genome-wide expression data have become a key task, including the aspect of feature selection. The basic scientific question of interest here for the lung cancer data is: which genes were most linked to the lung cancer?

Perhaps, the most popular way would be to apply a penalized regression method. For instance, Lasso selected 12 genes. However, the reliability of such results is a big issue, as mentioned already (see, e.g., Nan and Yang 2014). Two alternative approaches may be taken to address the question: via random forest importances and multiple hypothesis testing (Subramanian et al. 2005). As is pointed out in Subramanian et al. (2005), no genes are considered significantly related to the response at a 5% significance level by multiple hypothesis testing. From Table 6 (only top 5 are shown), random forest considers a number of genes to be more or less equally important, which does not seem to be very helpful in terms of telling the researcher if any gene(s) could be said to be far more important than the rest. In addition, the two random forest importance measures differ substantially in ranking of the genes. Thus, the two methods do not seem to reliably single out a few genes as most important to the lung cancer. Can SOIL bring some new insight?

Table 6. Top 5 variables for different variable importance measures of the Lung Cancer Data.

ARM	BIC-p	RFI1	RFI2
ENO2(0.999)	ENO2(0.235)	IL12RB1(2.383)	PAICS(0.222)
RHOG(0.086)	RHOG(0.0215)	UBE2C(2.188)	PSMA6(0.184)
PGAM1(0.005)	PGAM1(0.000)	EEF1A1(1.954)	RHOG(0.156)
MICB(0.002)	MICB(0.000)	DPF1(1.893)	IL12RB1(0.153)
DBP(0.001)	DBP(0.000)	P4HA1(1.883)	UBE2C(0.145)

We present two SOIL importances also in Table 6. SOIL-ARM views ENO2 absolutely important for the response, and SOIL-BIC-p also gives it an importance value much larger than all other genes (in this example, the BIC-p weighting seems too aggressive in pursuing parsimony, giving a large weight on the null model with intercept only). RHOG comes next, with importance values by SOIL-ARM/BIC-p much smaller than those of ENO2 but larger relative to the rest. Given the really small-sample size, RHOG might be potentially important should a larger sample size be used in a future study. We emphasize that SOIL importance is not meant to offer the final say, but it provides stable insight on which covariates are most important for explaining the response in the parametric modeling.

To further support the results of SOIL importances in Table 6, we carry out a cross-examination, in which the top two genes for SOIL-ARM (SOIL-BIC-p) and top five genes for RFI1(RFI2) are selected as the true variables, respectively (note that using more variables based on random forest gives even less reliable results for random forest). A Bernoulli distribution with probability $\hat{p}$ is used to generate the new response Y_{new}, where the estimated probabilities via logistic regression and vote proportion in random forest are used as the $\hat{p}$ for the home-game of SOIL and random forest, respectively. Figure 10 shows that the SOIL methods are self-consistent in the sense that it can identify the important variables in their home-game. Random forests are not self-consistent since the maximum variable importance of the unimportant variables is larger than those important ones. In the home-game of RFI1 and RFI2, SOIL does not recognize any true variables as important. The main reason is that the underlying generating process is nonparametric (with very weak signal), for which SOIL is not intended to be applicable. Overall, the SOIL importance measures seem to be well-supported in the multiple aspects above.

7. Conclusion and Discussion

Variable importance is aimed to find the important variables for explanation or prediction of the response. The motivation is most natural but the task of devising an importance measure is quite tricky. Several challenges immediately arrive: (1) importance depends on the goal of the analysis and application. Different goals may require different importance measures.

(2) Should importance be based on parametric models or nonparametric models? Both seem to be valuable in our view. (3) Should the importance measure be purely relative to compare different variables or should their values have some meaning on their own?

The topic is even controversial, with attitude ranging from enthusiasm in research and/or application, to reluctant acceptance as a practical approach to deal with many predictors, to total pessimism on the topic that dismisses the possibility of general successes. The different opinions are all valid, properly reflecting the complexity and multi-facet nature of the problem.

In our opinion, there are two important facts to keep in mind. One is that people crave for importance measures, love ranking, and they put them in use. This calls for more research on the topic. The other is that the currently dominating practice is still "winner-takes-all", which is definitely a culprit of irreproducibility of many research results. For reasonably complex data, making inference and decision based on a final selected model can lead to severely biased conclusions. A reliable importance measure can provide much needed complementary information to that from a final model and substantially improve the reliability of data analysis.

We have investigated the variable importance in linear regression and classification cases. The proposed new variable importance measure (SOIL) is driven by model combination for considering more than a single model, thus giving us an understanding of all the variables, instead of only the "important" ones in view of a single model. It is seen from both the simulation results and the real-data examples that the SOIL approach has several desirable features such as exclusion/inclusion, order preserving and robustness in several aspects, and performs very well compared to other variable importance measures considered.

As Grömping (2015) pointed out in her article, there is no commonly accepted theoretical framework in the variable importance area. Not surprisingly, many critiques on variable importance measures come up. Ehrenberg (1990) pointed out that one should focus on the underneath causal mechanism instead of the relative importance. We think SOIL is satisfactory in this regard. First, given enough information, SOIL assigns variable importance close to one for these true predictors, which is consistent with revealing the causal relationship between the response and the predictors. Second, the SOIL importance of a variable goes beyond relative assessment of the variables and it gives an absolute sense on how much a variable is needed in the linear modeling with the available information. In regression settings, data analysts often use t statistic or p-value to see if a variable is significant or not. Kruskal and Majors (1989) pointed out that this pertains to a different concept. In their view, variable importance is a population property, while significance is a property of both population and sample. To us, since all models are only approximations to model the data, there is advantage to treat variable importance measures as data-dependent quantities that reflect the nature of the data. SOIL intends to do just that.

Note that the two importance measures by the random forecast are not based on parametric modeling. When the GLM framework does not work for the data, our SOIL approach may not provide valuable information while random forest based ones may.

To be fair, it may be debatable if a variable that has some predictive power (one way or another) but is not needed in the best model should be given significant (reasonably strong) importance or not. Our view is that it seems rare to consider the covariates only individually and thus it is better to reflect the goal of finding the best set of covariates to explain the response in the importance measures. From this angle, while giving out relevant variables is certainly useful, it may not be most essential from a modeling perspective.

Through our simulation work, we have shown that the other methods often give clearly higher importance to variables that are not in the true model and/or give lower values for some variables in the true model when the covariates are correlated, error variance is large, or there are interaction terms. In real applications, these situations occur rather commonly. Thus, the results seem to suggest that when sparse modeling is the goal, those importance measures may not directly provide objective variable assessment information.

Appendix

Proof of Theorem 1.

Proof. Denote by $\mathcal{A}^*\backslash\mathcal{A}^k$ the set of variables contained in $\mathcal{A}^*$ but not in $\mathcal{A}^k$. Since

$$\begin{aligned}\frac{\sum_{k=1}^K w_k|\mathcal{A}^*\backslash\mathcal{A}^k|}{r^*} &= \frac{\sum_{k=1}^K w_k \sum_{j\in\mathcal{A}^*} I(j\notin\mathcal{A}^k)}{r^*}\\ &= \frac{\sum_{j\in\mathcal{A}^*}\sum_{k=1}^K w_k I(j\notin\mathcal{A}^k)}{r^*}\\ &= \frac{\sum_{j\in\mathcal{A}^*}\sum_{k=1}^K w_k (1-I(j\in\mathcal{A}^k))}{r^*}\\ &= \frac{\sum_{j\in\mathcal{A}^*}(1-S_j)}{r^*}.\end{aligned}$$

and by the definition of weak consistency,

$$0 \le \frac{\sum_{k=1}^K w_k|\mathcal{A}^*\backslash\mathcal{A}^k|}{r^*} \le \frac{\sum_{k=1}^K w_k|\mathcal{A}^k\nabla\mathcal{A}^*|}{r^*} \xrightarrow{p} 0.$$

Hence,

$$\frac{\sum_{j\in\mathcal{A}^*}(1-S_j)}{r^*} \xrightarrow{p} 0.$$

On the other hand,

$$\begin{aligned}\frac{\sum_{j\notin\mathcal{A}^*} S_j}{r^*} &= \frac{\sum_{j\notin\mathcal{A}^*}\sum_{k=1}^K w_k I(j\in\mathcal{A}^k)}{r^*}\\ &= \frac{\sum_{k=1}^K w_k \sum_{j\notin\mathcal{A}^*} I(j\in\mathcal{A}^k)}{r^*}\\ &= \frac{\sum_{k=1}^K w_k|\mathcal{A}^k\backslash\mathcal{A}^*|}{r^*}\\ &\le \frac{\sum_{k=1}^K w_k|\mathcal{A}^k\nabla\mathcal{A}^*|}{r^*} \xrightarrow{p} 0.\end{aligned}$$

□

Proof of Theorem 2.

Proof. Assume $\frac{|\overline{\mathcal{A}}_c|}{r^*}$ does not converge to 0 in probability as n tends to infinity (r^* may or may not depend on n), then there exists a positive constant ϵ_0, such that $P(\frac{|\overline{\mathcal{A}}_c|}{r^*} \geq \epsilon_0)$ does not converge to 0. On the other hand,

$$\begin{aligned}\frac{\sum_{j\in\mathcal{A}^*}(1-S_j)}{r^*} &= \frac{\sum_{j\in\mathcal{A}^*, S_j\leq c}(1-S_j)}{r^*} + \frac{\sum_{j\in\mathcal{A}^*, S_j>c}(1-S_j)}{r^*} \\ &\geq \frac{\sum_{j\in\mathcal{A}^*, S_j\leq c}(1-S_j)}{r^*} \\ &\geq \frac{\sum_{j\in\mathcal{A}^*, S_j\leq c}(1-c)}{r^*} \\ &= (1-c)\frac{\sum_{j\in\mathcal{A}^*} I(S_j\leq c)}{r^*} \\ &= (1-c)\frac{|\overline{\mathcal{A}}_c|}{r^*}.\end{aligned}$$

So we have $P(\frac{\sum_{j\in\mathcal{A}^*}(1-S_j)}{r^*} \geq (1-c)\epsilon_0) \geq P(\frac{|\overline{\mathcal{A}}_c|}{r^*} \geq \epsilon_0)$, which does not converge to 0. But this contradicts with Theorem 1. Hence, we have $\frac{|\overline{\mathcal{A}}_c|}{r^*} \xrightarrow{p} 0$. Similarly, we can prove $\frac{|\mathcal{A}_c|}{r^*} \xrightarrow{p} 0$. □

Supplementary Materials

R-package for SOIL: R-package `SOIL` containing code to compute the SOIL importance measure described in the article. (GNU zipped tar file)
Real datasets: Datasets BGS and Bardet used in the illustration of SOIL in Section 6. (.rda file)
Text document: Supplementary materials for "Sparsity Oriented Importance Learning for High-dimensional Linear Regression". (.pdf file)

Acknowledgment

The authors are grateful to the Editors and Referees for their constructive and insightful suggestions and comments, which improve the article in both organization and content.

Funding

Yang's research is partially supported by NSERC RGPIN-2016-05174.

References

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory* (Tsahkadsor, 1971), Akadémiai Kiadó, Budapest, pp. 267–281. [1799]
Archer, K. J., and Kimes, R. V. (2008), "Empirical Characterization of Random Forest Variable Importance Measures," *Computational Statistics & Data Analysis*, 52, 2249–2260. [1797]
Braun, M. T., and Oswald, F. L. (2011), "Exploratory Regression Analysis: A Tool for Selecting Models and Determining Predictor Importance," *Behavior Research Methods*, 43, 331–339. [1797]
Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1797,1801]
Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [1799]
Budescu, D. V. (1993), "Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression," *Psychological Bulletin*, 114, 542. [1797]
Chen, L., Giannakouros, P., and Yang, Y. (2007), "Model Combining in Factorial Data Analysis," *Journal of Statistical Planning and Inference*, 137, 2920–2934. [1798]
Cheng, T.-C. F., Ing, C.-K., and Yu, S.-H. (2015), "Toward Optimal Model Averaging in Regression Models with Time Series Errors," *Journal of Econometrics*, 189, 321–334. [1799]
Cheng, X., and Hansen, B. E. (2015), "Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach," *Journal of Econometrics*, 186, 280–293. [1799]
Chevan, A., and Sutherland, M. (1991), "Hierarchical Partitioning," *The American Statistician*, 45, 90–96. [1797]
Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. (2006), "Homozygosity Mapping with Snp Arrays Identifies TRIM32, an E3 Ubiquitin Ligase, as a Bardet–Biedl Syndrome Gene (BBS11)," *Proceedings of the National Academy of Sciences*, 103, 6287–6292. [1808]
Cook, R. D., and Weisberg, S. (2009), *Applied Regression Including Computing and Graphics*, Vol. 488, New York: Wiley. [1806]
Ehrenberg, A. S. C. (1990), "The Unimportance of Relative Importance," *The American Statistician*, 44, 260–260. [1810]
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. [1799,1800]
Feldman, B. (2000), "The Proportional Value of a Cooperative Game," in *Econometric Society World Congress 2000 Contributed Papers* (No. 1140). Econometric Society. [1797]
——— (2005), "Relative Importance and Value," *Available at SSRN 2255827*. [1797]
Ferrari, D., and Yang, Y. (2015), "Confidence Sets for Model Selection by *F*-testing," *Statistica Sinica*, 25, 1637–1658. [1797]
Grömping, U. (2015), "Variable Importance in Regression Models," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 137–152. [1797,1804,1810]
Grömping, U., et al. (2006), "Relative Importance for Linear Regression in R: The Package Relaimpo," *Journal of Statistical Software*, 17, 1–27. [1801]
Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [1799]
Hapfelmeier, A., Hothorn, T., Ulm, K., and Strobl, C. (2014), "A New Variable Importance Measure for Random Forests with Missing Data," *Statistics and Computing*, 24, 21–34. [1797]
Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 98, 879–899. [1799]
Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401. [1799]
Hoffman, P. J. (1960), "The Paramorphic Representation of Clinical Judgment," *Psychological Bulletin*, 57, 116. [1797]
Huang, J., Ma, S., and Zhang, C.-H. (2008), "Adaptive Lasso for Sparse High-Dimensional Regression Models," *Statistica Sinica*, 18, 1603–1618. [1808]
Ioannidis, J. P., and Khoury, M. J. (2011), "Improving Validation Practices in "omics" Research," *Science*, 334, 1230–1232. [1798]
Ishwaran, H. (2007), "Variable Importance in Binary Regression Trees and Forests," *Electronic Journal of Statistics*, 1, 519–537. [1797]
Kruskal, W., and Majors, R. (1989), "Concepts of Relative Importance in Recent Scientific Literature," *The American Statistician*, 43, 2–6. [1810]
Lai, R. C., Hannig, J., and Lee, T. C. (2015), "Generalized Fiducial Inference for Ultrahigh-Dimensional Regression," *Journal of the American Statistical Association*, 110, 760–772. [1801]
Leung, G., and Barron, A. R. (2006), "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52, 3396–3410. [1799]
Li, K.-C., Lue, H.-H., and Chen, C.-H. (2000), "Interactive Tree-Structured Regression Via Principal Hessian Directions," *Journal of the American Statistical Association*, 95, 547–560. [1806]
Liang, H., Zou, G., Wan, A. T., and Zhang, X. (2011), "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 106, 1053–1066. [1799]

Liaw, A., and Wiener, M. (2002), "Classification and Regression by RandomForest," *R News*, 2, 18–22. [1801]

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980), *Introduction to Bivariate and Multivariate Analysis.*, number 519.535 L743, Scott, Foresman. [1797,1801]

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013, December), "Understanding Variable Importances in Forests of Randomized Trees," in *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pp. 431–439, New York: Curran Associates Inc. [1797]

McNutt, M. (2014), "Raising the Bar," *Science*, 345, 9. [1798]

Meinshausen, N., and Bühlmann, P. (2010), "Stability Selection," *Journal of the Royal Statistical Society*, Series B, 72, 417–473. [1806]

Nan, Y., and Yang, Y. (2014), "Variable Selection Diagnostics Measures for High-Dimensional Regression," *Journal of Computational and Graphical Statistics*, 23, 636–656. [1798,1800,1809]

Rolling, C. A., and Yang, Y. (2014), "Model Selection for Estimating Treatment Effects," *Journal of the Royal Statistical Society*, Series B, 76, 749–769. [1807]

Schwarz, G., et al. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [1799]

Stodden, V. (2015), "Reproducing Statistical Results," *Annual Review of Statistics and Its Application*, 2, 1–19. [1798]

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008), "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, 9, 1. [1797]

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007), "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, 8, 25. [1797]

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005), "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences*, 102, 15545–15550. [1809]

Theil, H., and Chung, C. (1988), "Information-Theoretic Measures of Fit for Univariate and Multivariate Linear Regressions," *The American Statistician*, 42, 249–252. [1797]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1799,1800]

Tuddenham, R. D., and Snyder, M. M. (1954), "Physical Growth of California Boys and Girls from Birth to Eighteen Years," *Publications in Child Development*, 1, 183. [1806]

Yang, Y. (2000), "Adaptive Estimation in Pattern Recognition by Combining Different Procedures," *Statistica Sinica*, 10, 1069–1090. [1799]

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574–588. [1799]

Yang, Y. (2003), "Regression with Multiple Candidate Models: Selecting or Mixing?," *Statistica Sinica*, 13, 783–809. [1798]

Yang, Y. (2005), "Can the Strengths of AIC and BIC be Shared? A Conflict between Model Indentification and Regression Estimation," *Biometrika*, 92, 937–950. [1798]

Yang, Y. (2007), "Consistency of Cross Validation for Comparing Regression Procedures," *The Annals of Statistics*, 35, 2450–2473. [1799]

Yang, Y., and Barron, A. R. (1998), "An Asymptotic Property of Model Selection Criteria," *IEEE Transactions on Information Theory*, 44, 95–116. [1800]

Yuan, Z., and Ghosh, D. (2008), "Combining Multiple Biomarker Models in Logistic Regression," *Biometrics*, 64, 431–439. [1799]

Zhang, C. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [1799,1800]

Zhang, X., Lu, Z., and Zou, G. (2013), "Adaptively Combined Forecasting for Discrete Response Time Series," *Journal of Econometrics*, 176, 80–91. [1799]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1799]

Supplemental Materials for "Sparsity Oriented Importance Learning for High-dimensional Linear Regression"

Chenglong Ye [*], Yi Yang [†], Yuhong Yang [‡]

August 27, 2017

Part A: Weighting using generalized fiducial inference.

Based on Fisher's controversial fiducial idea, Lai et al. (2015) proposed the generalized fiducial inference applied to "large p small n" problem. Their paper concerns the generalized fiducial inference for the linear regression case. For each candidate model $\mathcal{A}^k$, the fiducial probability for the model is

$$p(\mathcal{A}^k) \propto R(\mathcal{A}^k) \equiv \Gamma(\frac{n-\left|\mathcal{A}^k\right|}{2})(\pi RSS_{\mathcal{A}^k})^{-\frac{n-\left|\mathcal{A}^k\right|-1}{2}} n^{-\frac{\left|\mathcal{A}^k\right|+1}{2}} \left(\begin{array}{c} p \\ \left|\mathcal{A}^k\right| \end{array} \right)^{-\gamma},$$

where $RSS_{\mathcal{A}^k}$ is the residual sum of squares of $\mathcal{A}^k$. For a practical reason, the authors approximate the above fiducial probability by

$$r(\mathcal{A}^k) \approx R(\mathcal{A}^k) / \sum_{l=1}^{K} R(\mathcal{A}^l).$$

We can use $r(\mathcal{A}^k)$ as the weight w_k for each candidate model. It is shown in their paper that the true model will have the highest fiducial probability among all the candidate models.

[*]School of Statistics, University of Minnesota (yexxx323@umn.edu)

[†]Corresponding author, Department of Mathematics and Statistics, McGill University (yi.yang6@mcgill.ca)

[‡]School of Statistics, University of Minnesota (yangx374@umn.edu)

Part B: Additional simulation results.

In this part, we provide the results of Example S1-S6, whose settings are described in Table 1 of the main body of the article. These results support our conclusions as discussed in Section 5.1.

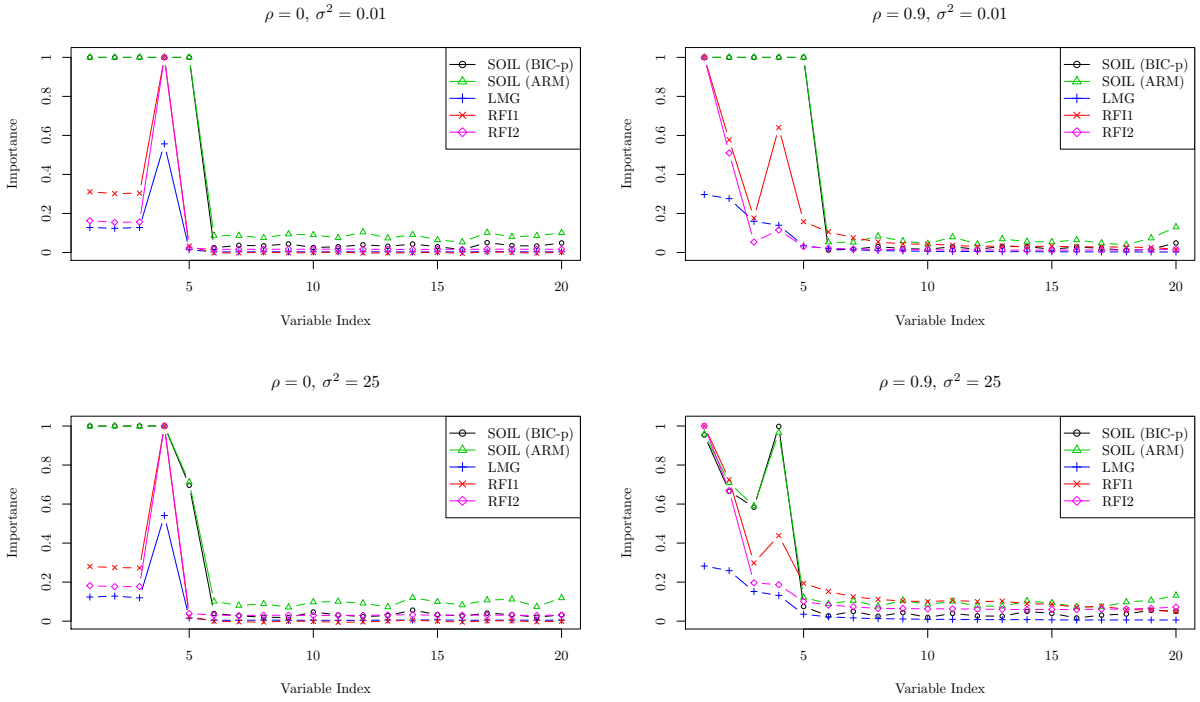

Figure S1: Simulation results for Example S1, where $n = 150$, $p = 20$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)$.

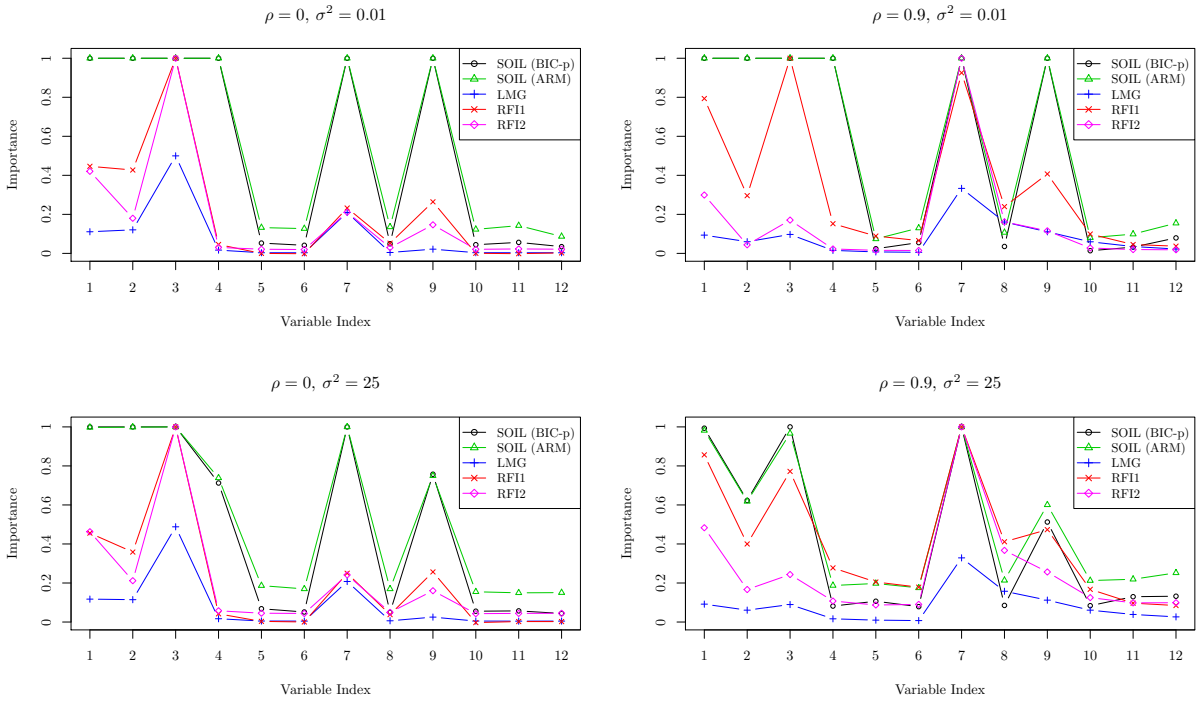

Figure S2: Simulation results for Example S2, where $n = 150$, $p = 6$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, -6\sqrt{2}, \frac{3}{4}, 0, 0)^\intercal$. Add $(X_1^2, X_2^2, X_3^2, X_4^2, X_5^2, X_6^2)$ and corresponding coefficients $(\beta_7^*, \beta_8^*, \ldots, \beta_{12}^*)^\intercal = (4, 0, 1, 0, 0, 0)^\intercal$.

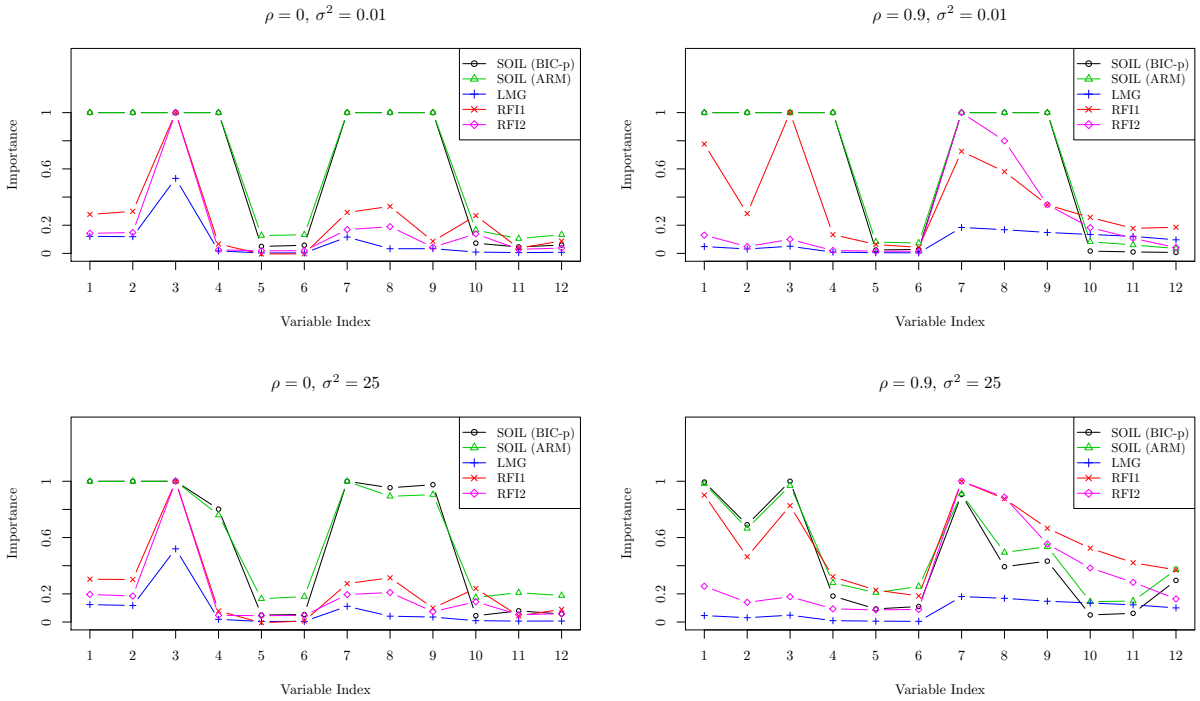

Figure S3: Simulation results for Example S3, where $n = 150$, $p = 6$. The true coefficient $\boldsymbol{\beta}^* = (4, 4, -6\sqrt{2}, \frac{3}{4}, 0, 0)^\intercal$. Add $(X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4)$ and corresponding coefficients $(\beta_7^*, \beta_8^*, \ldots, \beta_{12}^*)^\intercal = (4, 2, 2, 0, 0, 0)^\intercal$.

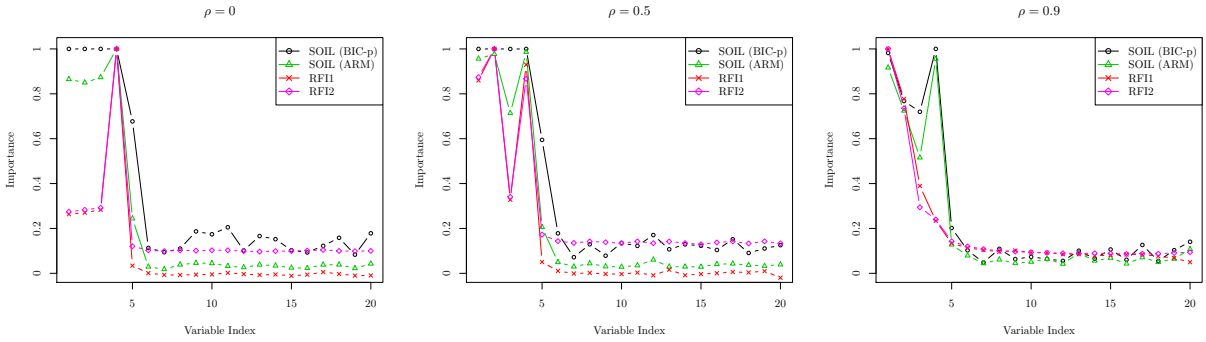

Figure S4: Simulation results for Example S4, where $n = 150$, $p = 20$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)$.

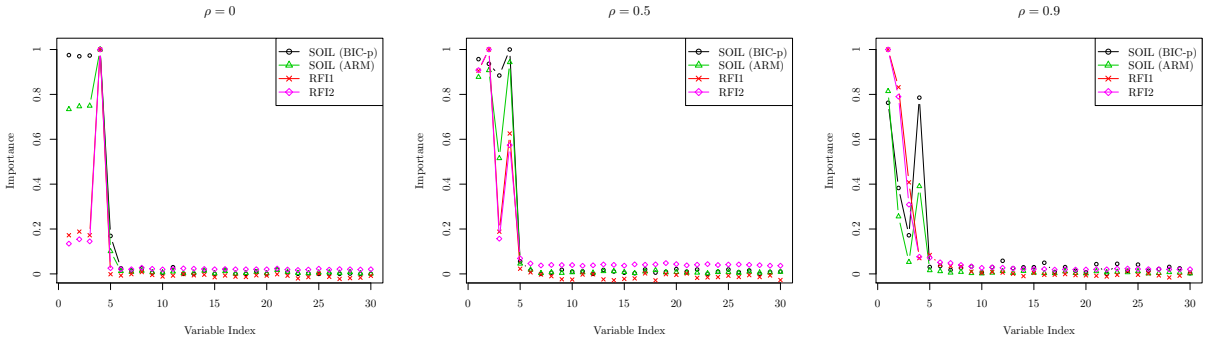

Figure S5: Simulation results for Example S5, where $n = 100$, $p = 200$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)$.

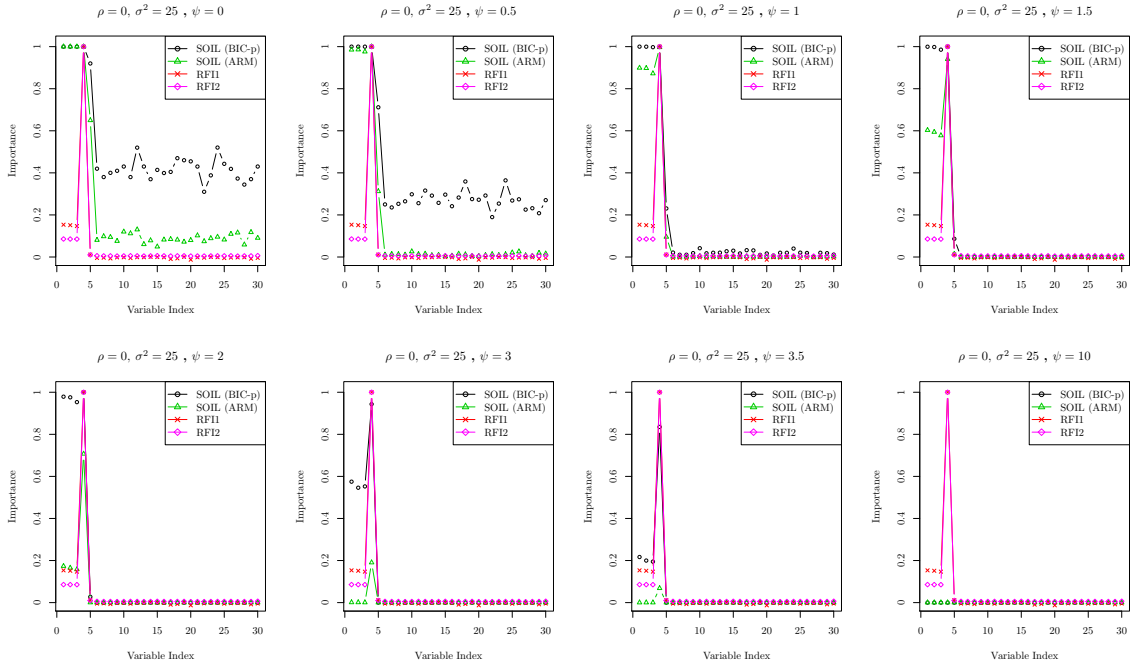

Figure S6: Sensitivity analysis of ψ, where $n = 100$, $p = 200$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)$.

Part C: Comparison with stability selection.

In this subsection, we present a comparison of SS (Meinshausen & Bühlmann 2010) importance and our SOIL importance.

The simulation data $\{y_i, \mathbf{x}_i\}_{i=1}^n$ is generated from the linear model $y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \epsilon_i$, $\epsilon \sim N(0, \sigma^2)$. We generate $\mathbf{x}_i$ from multivariate normal distribution $N_p(0, \Sigma)$. For each element

Σ_{ij} of Σ, $\Sigma_{ij} = \rho^{|i-j|}$, i.e. the correlation of X_i and X_j is $\rho^{|i-j|}$. We consider two cases, the settings of which are listed in Table S1.

Example	n	p	ρ	σ^2	Coefficients
1	100	20	0	0.01	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)^\intercal$
2	100	20	0.7	0.1	$\boldsymbol{\beta}^* = (4, 0, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)^\intercal$

Table S1: Simulation settings for SS

It can be seen from Tables S2 and S3 that SS does not give enough importance to the true variable X_5 in Example 1 while it more strongly supports the noise variable X_2 than the true variable X_5 in Example 2, which leads to unavoidable incorrect variable selection regardless of the cutoff to be used to decide if a variable is in or out based on its importance. In contrast, SOIL-ARM and SOIL-BIC-p pick all the important variables and leave noise variables out. From these results, together with the fact that the main goal of SS is not on variable importance, we have not considered stability selection in the main simulations in this work.

Method/Variable	X_1	X_2	X_3	X_4	X_5	max of rest
SOIL-ARM	1.00	1.00	1.00	1.00	1.00	0.12
SOIL-BIC-p	1.00	1.00	1.00	1.00	1.00	0.07
Stability Selection	0.99	0.99	0.99	1.00	0.02	0.002

Table S2: Variable importance for Example 1.

Method/Variable	X_1	X_2	X_3	X_4	X_5	max of rest
SOIL-ARM	1.00	0.15	1.00	1.00	1.00	0.14
SOIL-BIC-p	1.00	0.06	1.00	1.00	1.00	0.05
Stability Selection	1.00	0.44	0.94	1.00	0.26	0.05

Table S3: Variable importance for Example 2.

Part D: Stability comparison of SOIL and Lasso.

We conduct a stability comparison of our methods and Lasso at a reduced sample size to show that our method is more stable than Lasso against small changes in the data. The simulation data $\{y_i, \mathbf{x}_i\}_{i=1}^n$ is generated from the linear model $y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}^* + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma^2 = 0.01$. $\mathbf{x}_i$ is generated from $N_p(0, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. We set $n = 50$, $p = 200$ and

$\boldsymbol{\beta}^* = (4, 4, -6\sqrt{2}, 4/3, 0, 0, 4, 0, 1, 0, \ldots, 0)^\intercal$. We randomly remove 10 observations from the dataset and use the remaining data to compute the corresponding SOIL-BIC-p importances and the Lasso coefficients. The results are recorded over 100 replications and shown in Figure S7. We can see that, for each run with the reduced sample size, the result for the SOIL importance is pretty consistent, while the result for the Lasso coefficients varies considerably, indicating that the SOIL importance has the continuity property with respect to a reduced sample size and is more stable than Lasso.

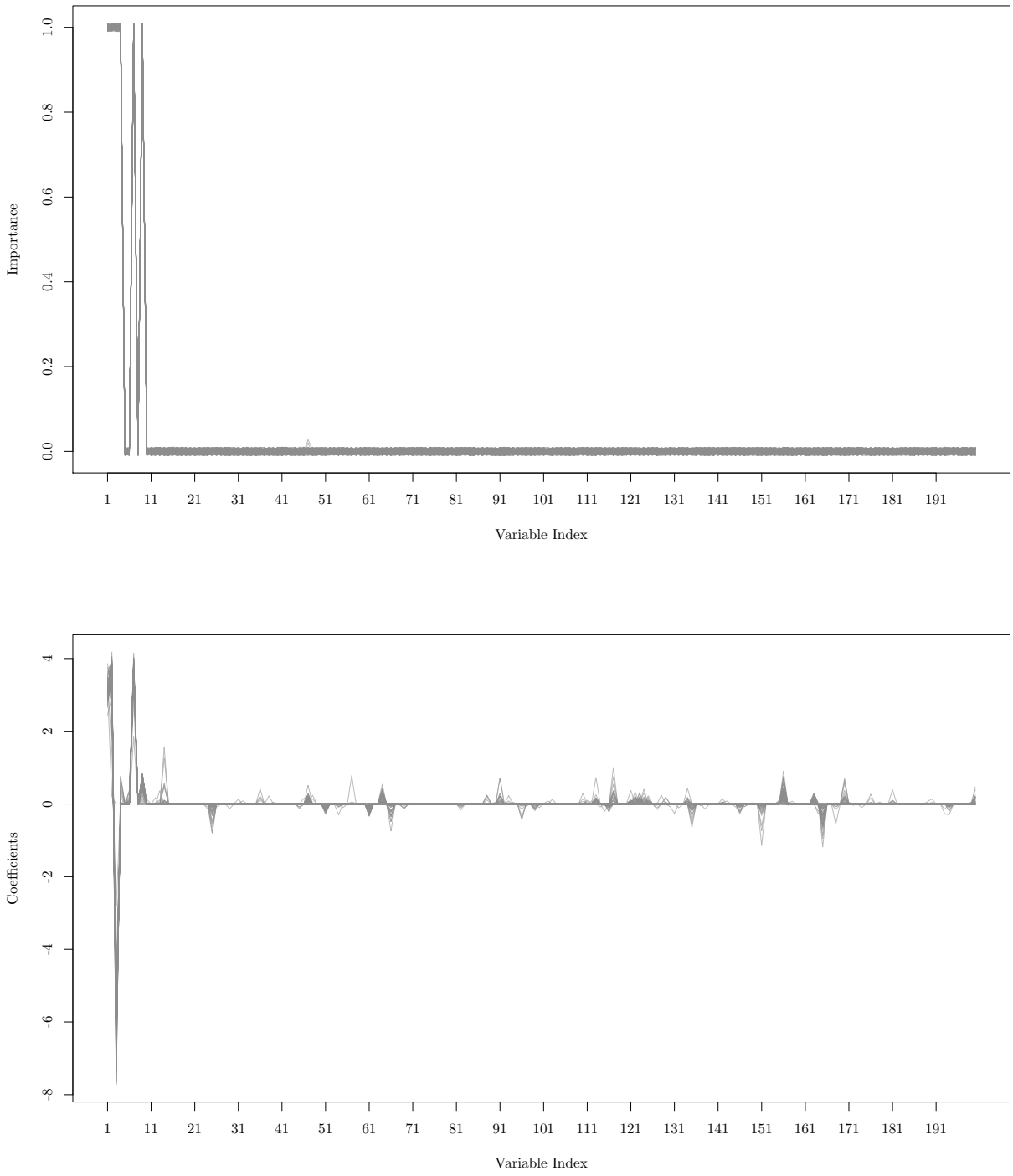

Figure S7: Stability comparison of SOIL-BIC-p and Lasso at a reduced sample size for 100 replications. Top panel: SOIL-BIC-p importances. Bottom panel: Lasso coefficients. Each grey line represents the result from one replication.

References

Lai, R. C., Hannig, J. & Lee, T. C. (2015), 'Generalized fiducial inference for ultrahigh-dimensional regression', *Journal of the American Statistical Association* **110**(510), 760–772.

Meinshausen, N. & Bühlmann, P. (2010), 'Stability selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4), 417–473.