Model Selection and Estimation for High-dimensional Data Analysis

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Chenglong Ye

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Yuhong Yang, Adviser

June 2019

Acknowledgements

I'd like to thank my advisor, Professor Yuhong Yang, for his patient guidance and utmost kindness as a supervisor, enthusiasm as a researcher, sharp insights and rigorousness as a scholar, during my Ph.D. studies. This thesis couldn't have been done without his guidance of both the research and the writing. A thousand words in one sentence: I am lucky and so proud of being his student.

I would like to thank my committee members, Professors Hui Zou, Professor Weihua Guan and Professor Lan Liu for their encouraging comments and valuable questions. I would also like to thank my collaborators Charles Doss, Jie Ding and Jueyu Wang for research opportunities and inspiring discussions to work on interesting and challenging projects. I thank all my friends for making my PhD life full of happiness.

Last but not least, I thank my parents, Yongqing Ye and Xiufei Hong, for their meticulous parental guidance and support for my pursuit of PhD.

Abstract

In the era of big data, uncovering useful information and hidden patterns in the data is prevalent in different fields. However, it is challenging to effectively select input variables in data and estimate their effects. In this thesis, our goal is to develop reproducible statistical approaches that provide mechanistic explanations of the phenomenon observed in big data analysis. The thesis contains two parts: variable selection and model estimation. The first part investigates how to measure and interpret the usefulness of an input variable using an approach called "variable importance learning" and builds tools (methodology and software) that can be widely applied. We propose two variable importance measures, a parametric measure SOIL and a nonparametric measure CVIL, using the idea of model combining and cross validation respectively. 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. The CVIL method possesses desirable theoretical properties and enhance the interpretability of many mysterious but effective machine learning methods. The second part focuses on how to estimate the effect of a useful input variable in the case where interaction of two input variables exists. We investigate the minimax rate of convergence for regression estimation in high-dimensional sparse linear models with two-way interactions, and construct an adaptive estimator that achieves the minimax rate of convergence regardless of the true heredity condition and the sparsity indices.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the rapid development of technology, massive amounts of information/data are produced every day in this era of big data. In various fields such as engineering, computer science and finance, many statistical and machine learning methods are applied to uncover useful information and patterns behind these enormous datasets. However, the theoretical mechanisms of many predictive machine learning methods are not fully understood or even not understood at all. Developing reproducible statistical approaches that provide mechanistic explanations of the phenomenon observed in big data analysis is challenging and important.

In high-dimensional data analysis, a common problem is to select the relevant or important variables among an an enormous number of variables, followed by the problem of estimating the effects of these selected variables. Such problems are challenging and meanwhile of great importance in real life applications as following. For example, one lung cancer study (Subramanian et al., 2005) investigates the genetic mutations on critical genes that are related to lung cancer. A record is kept of the health status, which is classified as "good" or "poor", of 62 patients along with measurements of the activity of 5217 genes for each patient. To find out which genes are most related to lung cancer, the researchers may determine the variable importance of each gene and then investigate those genes with high variable importance. Discovering biologically relevant genes is essential in both early detection and treatment of such deadly diseases. For some data, it is crucial to consider the interaction

between two input variables. For example, strong interactions of two antigenic sites are observed during virus evolution (Han et al., 2016). This gene-gene interaction is a common component in disease analysis. When interaction exists, structures are usually required to describe the relationship, such as strong heredity. Strong heredity means that if the interaction of two input variables is included in a model, then both variables must individually be included in the model. Returning to the virus example, since the interaction of two antigenic sites is strongly related to virus evolution, each antigenic site should be considered relevant.

The thesis consists of two parts. In the first part, we investigate how to measure and interpret the usefulness of an input variable using an approach called "variable importance learning" and builds tools (methodology and software) that can be widely applied. We propose two variable importance measures, a parametric measure SOIL in Chapter 2 and a nonparametric measure CVIL in Chapter 3, corresponding to and respectively.

In Chapter 2, 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. The proofs of the results are in Appendix A. This paper corresponds to Ye et al. (2018).

In Chapter 3, we propose Cross Validation Importance Learning (CVIL), which can be applied to any parametric or nonparametric methods to help demystify how

these methods employ the input variables to make predictions. Given any specific method, by deleting a variable in the data set or replacing the variable with a constant, CVIL measures the relative difference of the predictive performance of the model from a cross-validation perspective. Under some mild conditions, CVIL is consistent in the sense that it converges to the theoretical variable importance as the sample size grows. Confidence intervals are constructed to show the reliability of the proposed CVIL importance measure. By simulations and real data examples, we show that CVIL provides a rank of variable importance attached to any seemingly uninterpretable predictive algorithm such as deep neural network.

The second part of the thesis focuses on how to estimate the effect of a useful input variable in the case where interaction of two input variables exists.

Chapter 4 corresponds to Ye and Yang (2019). In this chapter, we first investigate the minimax rate of convergence for regression estimation in high-dimensional sparse linear models with two-way interactions. We derive matching upper and lower bounds under three types of heredity conditions: strong heredity, weak heredity and no heredity. From the results: (i) A stronger heredity condition may or may not drastically improve the minimax rate of convergence. In fact, in some situations, the minimax rates of convergence are the same under all three heredity conditions; (ii) The minimax rate of convergence is determined by the maximum of the total price of estimating the main effects and that of estimating the interaction effects, which goes beyond purely comparing the order of the number of non-zero main effects r_1 and non-zero interaction effects r_2; (iii) Under any of the three heredity conditions, the estimation of the interaction terms may be the dominant part in determining the rate of convergence for two different reasons: 1) there exist more interaction terms than main effect terms or 2) a large ambient dimension makes it more challenging to estimate even a small number of interaction terms. Second, we construct an adaptive estimator that achieves the minimax rate of convergence regardless of the true heredity condition and the sparsity indices r_1,r_2.

Chapter 2

Sparsity Oriented Importance Learning for High-dimensional Linear Regression

2.1 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); 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, e.g., 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 et al. (1980) used the improved explained variance averaged over all possible orderings of predictors to provide a ranking of the predictors. Feldman et al. (1999) 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. (2008a) proposed conditional variable importance for random forest to correct the bias of variable importance when there exist correlated variables. Ferrari and Yang (2015) assess variable importance from a variable selection confidence set (VSCS) perspective.

In this Chapter, 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 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 Chapter, 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 et al. (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 Chapter is organized as follows. In Section 2.2, we introduce the proposed SOIL methodology and provide a theoretical understanding on some key aspects. Sections 2.3 and 2.4 present the details of choosing the candidate models and the weighting for SOIL in practice. In Section 2.5, 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 2.6.

2.2 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 2.4.

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 $\mathcal{A} = \{\mathcal{A}_k\}_{k=1}^K$, 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, 2010a) 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 $\mathcal{A}$ and the weighting $\mathbf{w}$, we define the SOIL importance measure for the j-th variable, $j \in \{1, \ldots, p\}$, as the accumulated sum of weights of the candidate models $\mathcal{A}^k$ that contains the j-th variable. That is

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

2.2.1 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

$$\frac{\sum_{k=1}^{K} w_k |\mathcal{A}^k \nabla \mathcal{A}^*|}{r^*} \xrightarrow{p} 0, \qquad \text{as } n \to \infty, \tag{2.1}$$

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

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

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 ((2.1)) 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 $\mathcal{A}$, 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 $\mathcal{A}$. 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 $\mathcal{A}$. For example, Buckland et al. (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 et al., 1997; Hansen, 2007; Liang et al., 2012; Cheng et al., 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); Zhang et al. (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:

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

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 $\overline{\mathcal{A}}_c = \{j \in \mathcal{A}^* : S_j \le c,\ j = 1,...,p\}$, $\underline{\mathcal{A}}_c = \{j \notin \mathcal{A}^* : S_j > c, j = 1,...,p\}$, then if $\mathbf{w}$ is weakly consistent, we have

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

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.

2.3 Implementation

2.3.1 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, i.e.

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

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

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(\beta_j), \tag{2.2}$$

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 ((2.2)), and nonconvex penalties including the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001)

$$\begin{aligned} p_\lambda(u) &= \lambda |u| I(|u| \leq \lambda) + \left\{ \lambda |u| - \frac{(\lambda - |u|)^2}{2(\gamma - 1)} \right\} I(\lambda < |u| \leq \gamma\lambda) \\ &\quad + \frac{(\gamma + 1)\lambda^2}{2} I(|u| > \gamma\lambda), \qquad (\gamma > 2), \end{aligned}$$

or the minimax concave penalty (MCP, Zhang (2010a))

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

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

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

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 $\mathcal{A}_{\text{SCAD}}$ 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 $\mathcal{A} = \{\mathcal{A}_{\text{Lasso}}, \mathcal{A}_{\text{AdaptiveLasso}}, \mathcal{A}_{\text{SCAD}}, \mathcal{A}_{\text{MCP}}\}$. The

individual penalized methods for producing $\mathcal{A}$ 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 $\mathcal{A}$ is enhanced.

2.3.2 Weighting

In this Chapter, 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 sub-vector 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 $\mathcal{A}^k \in \mathcal{A}$, 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:

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

  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 non-constant predictors. Then weight w_k for model $\mathcal{A}^k \in \mathcal{A}$ is computed by

$$w_k = \exp(-\frac{I_k}{2} - \psi C_k)/\sum_{l=1}^{K}\exp(-\frac{I_l}{2} - \psi C_l). \tag{2.3}$$

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 et al., 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 et al., 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 $\mathcal{A}$ is consistent, as long as at least one method is path-consistent, we have

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

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

2.3.3 Software

We provide our implementation of the SOIL importance measure in an official **R** package `SOIL`, which is publicly available on CRAN.

2.4 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

$$\Pr(Y=1|X=\mathbf{x}) = 1-\Pr(Y=0|X=\mathbf{x}) = \frac{e^{\mathbf{x}^\intercal\boldsymbol{\beta}^*}}{1+e^{\mathbf{x}^\intercal\boldsymbol{\beta}^*}}, \tag{2.4}$$

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 $\mathcal{A}^k \in \mathcal{A}$, 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:

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

- For each $\mathcal{A}^k$, compute the predicted probability $\widehat{p}^{(k)}(\mathbf{x}_{s,i}^{(k)})$ on the test set $\{i | i \in \mathbf{D}_2\}$.
- 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 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 $w_k = \frac{1}{L} \sum_{l=1}^{L} w_k^{(l)}$.

2.4.1 Weighting using information criteria with nonuniform priors

Similarly, the weight w_k for model $\mathcal{A}^k \in \mathcal{A}$ using BIC-p the information criterion can be computed in the same way as in ((2.3)) where $I_k^{\text{BIC}} = -2 \log \ell_k + 2 s_k \log n$, with $s_k = |\mathcal{A}^k|$ and ℓ_k being the maximized likelihood function for the logistic model $\mathcal{A}^k$.

2.5 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 et al., 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 $\mathcal{A}_{lasso}$, $\mathcal{A}_{SCAD}$ and $\mathcal{A}_{MCP}$ 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 $\mathcal{A}_{lasso}$, $\mathcal{A}_{SCAD}$ and $\mathcal{A}_{MCP}$ as our candidate set $\mathcal{A}$.

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

Model 2: Binomial. The i.i.d. 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 2.1 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 2.1– 2.6 and Figures A.1– A.6. Due to page restrictions, the figures of Example A.1– A.6 are only provided in the supplementary materials, while the summary of all the examples are discussed in the main part of the Chapter. 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 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 A.6 in the Supplementary Materials. We tried eight different values, i.e. $\psi \in \{0, 0.5, 1, 1.5, 2, 3, 3.5, 10\}$ 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, i.e. 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.

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	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$. 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	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$
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	7	$\boldsymbol{\beta}^* = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0\right)^\intercal$
6	5000	7	$\boldsymbol{\beta}^* = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0\right)^\intercal$
S4	150	20	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$
S5	100	200	$\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)^\intercal$

Table 2.1: Simulation settings

2.5.1 Relative performances of importance measures in several key aspects

A summary of the relevant properties of different important measures is provided in Table 2.2. 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.

	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	✓	✓	✓		✓

Table 2.2: 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.

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 A.1, 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 2.2 the noise variable X_{15} confuses LMG, RFI1 and RFI2. In Figure A.2 when $\rho = 0.9$, LMG, RFI1 and RFI2 assign relatively high values on the noise variable X_8. In Figure A.3 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 2.5 shows that in a binomial model with the decreasing coefficient vector $\boldsymbol{\beta}^* = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0\right)^\intercal$, 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 2.6 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 2.5 and 2.6, which differ only in the sample size, as shown in Figure 2.5 and Figure 2.6, 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 2.1, 2.2 and Figures A.1– A.5 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.

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 2.2, 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 2.2 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 A.2 and A.3 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 A.2 and A.3 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 2.3 and 2.4 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 2.3 have very similar

LMG and RFI2 values to those in Example 2.4. And RFI1 behaves wildly as it assigns very much different importances to the variables in the independence case ($\rho = 0$) of Example 2.3. 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 2.2. RFI1 and RFI2 do not preserve the order when correlation $\rho = 0.9$ and/or noise σ^2 is large.

High-dimensionality. SOIL-BIC-p, SOIL-ARM, RFI1 and RFI2 can work for high-dimensional data when $p > n$ as shown in Figure 2.1 and A.5. 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.

Non-negativity. SOIL-BIC-p, SOIL-ARM, LMG and IMG2 always yield non-negative importance value. However, RFI1 does not satisfy this criterion.

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

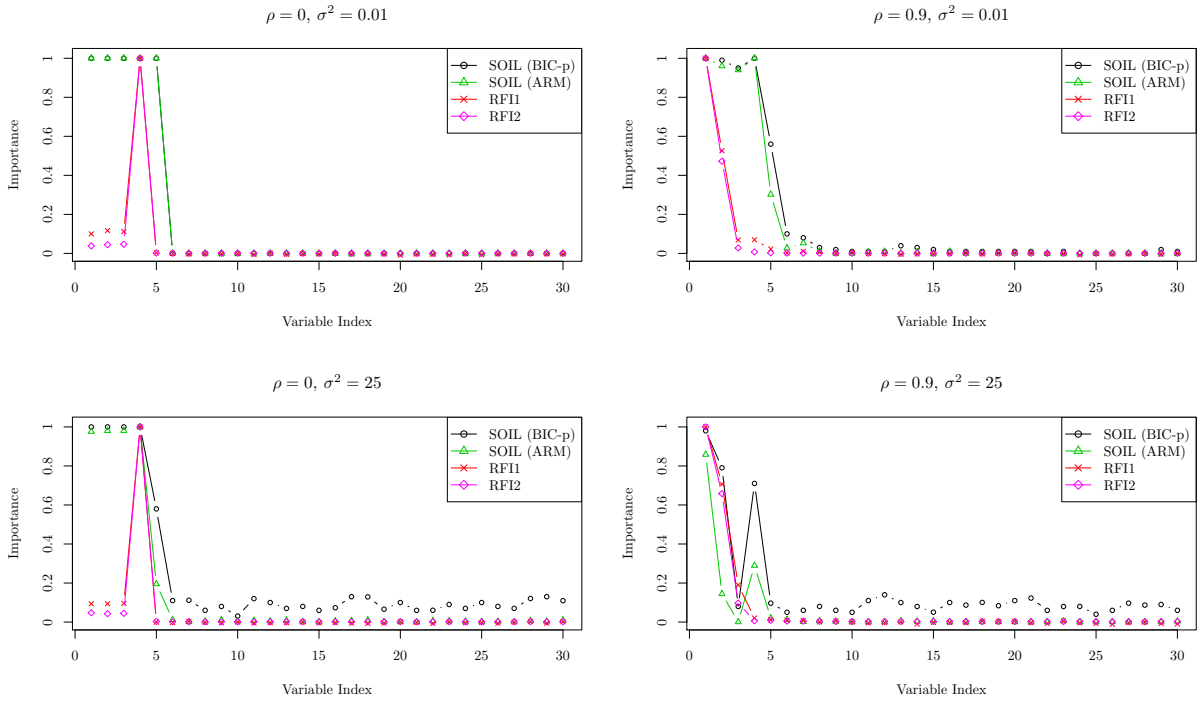


Figure 2.1: Simulation results for Example 1, where $n = 100$, $p = 1000$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)$.

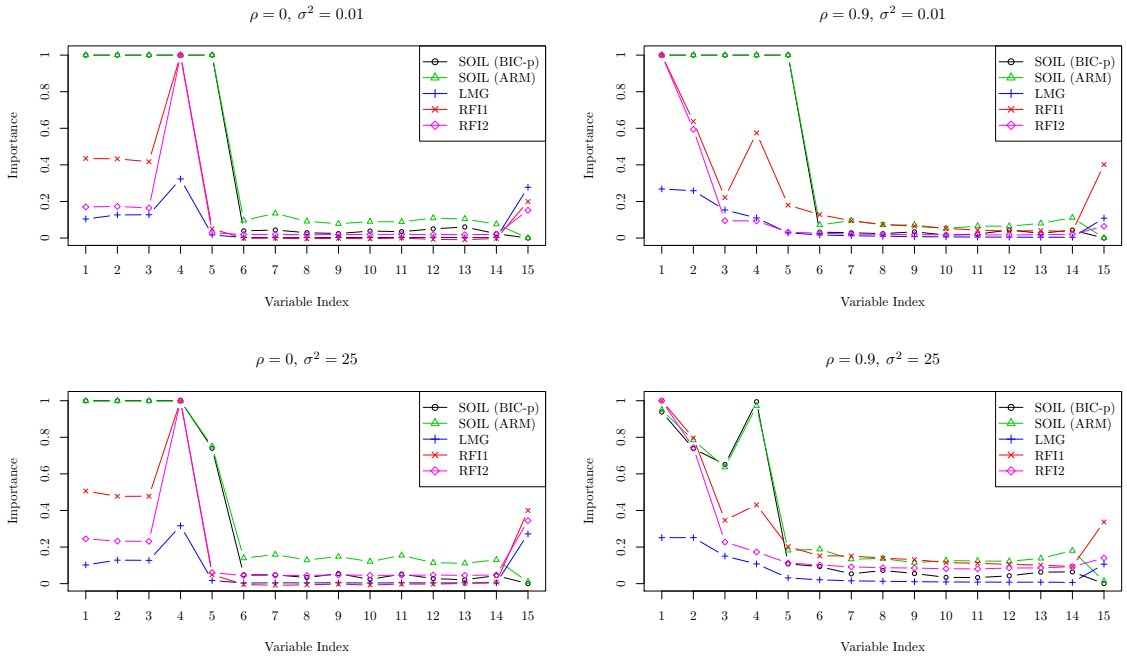


Figure 2.2: Simulation results for Example 2, where $n = 150$, $p = 14$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{4}{3}, 0, ..., 0)$. Add $X_{15} = 0.5 * X_1 + 2 * X_4 + e$ and corresponding $\beta^*_{15} = 0$, where $e \sim N(0, \sigma_e^2)$.

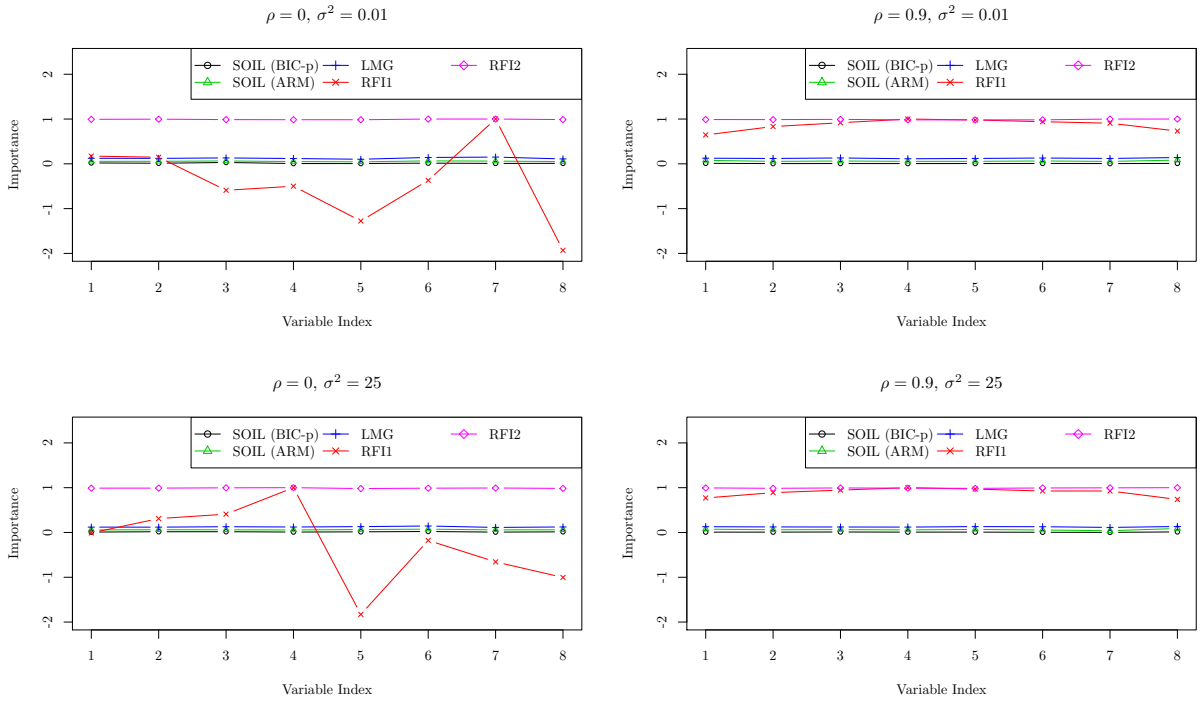


Figure 2.3: Simulation results for Example 3, where $n = 150$, $p = 8$. The true coefficients $\boldsymbol{\beta}^* = (0, \ldots, 0)^\intercal$.

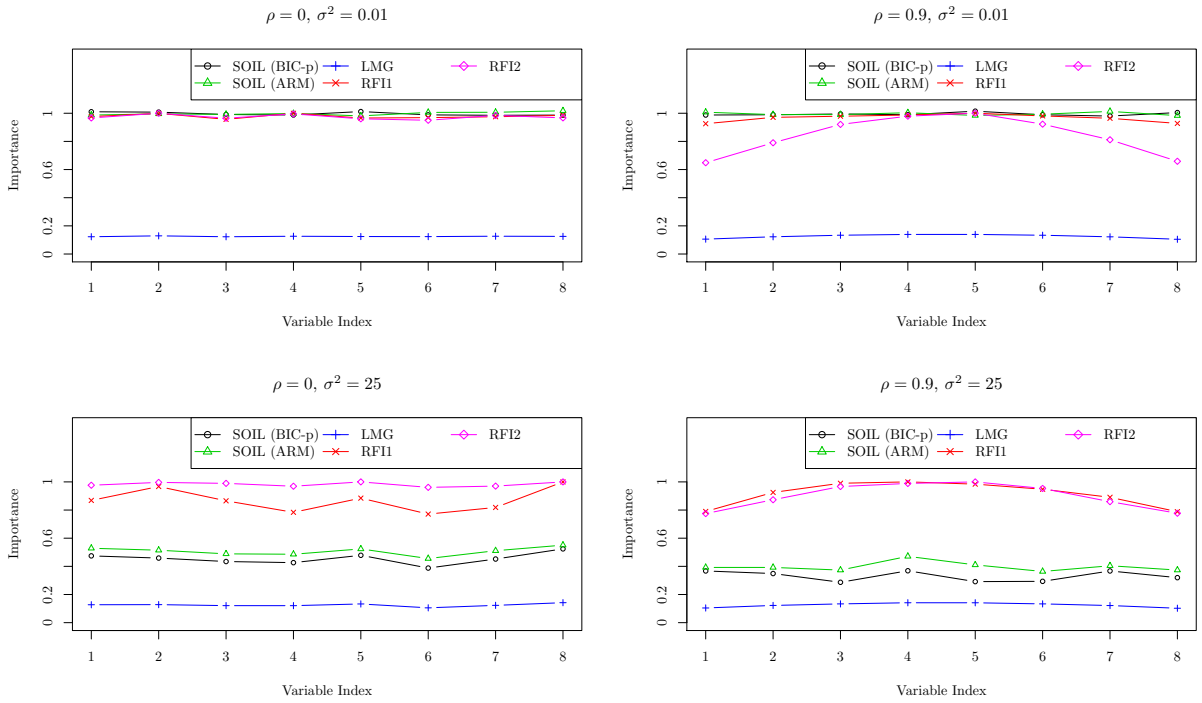


Figure 2.4: Simulation results for Example 4, where $n = 150$, $p = 8$. The true coefficients $\boldsymbol{\beta}^* = (1, \ldots, 1)^\intercal$.

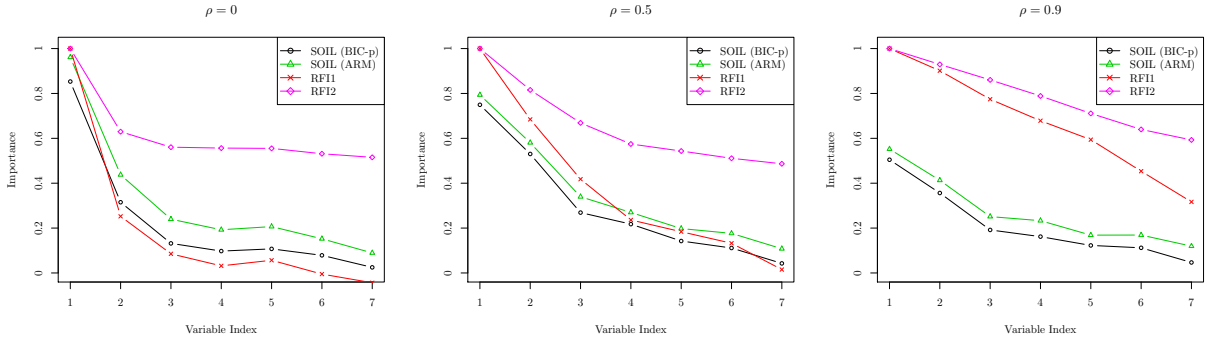


Figure 2.5: Simulation results for Example 5, where $n = 80$, $p = 6$. The true coefficients $\boldsymbol{\beta}^* = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0\right)^\intercal$.

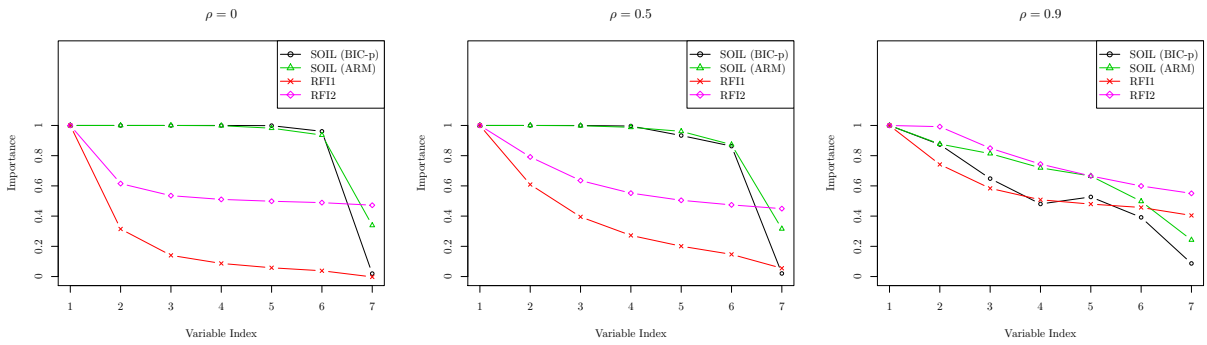


Figure 2.6: Simulation results for Example 6, where $n = 5000$, $p = 6$. The true coefficients $\boldsymbol{\beta}^* = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, 0\right)^\intercal$.

2.5.2 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.

2.5.3 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 k-th 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)

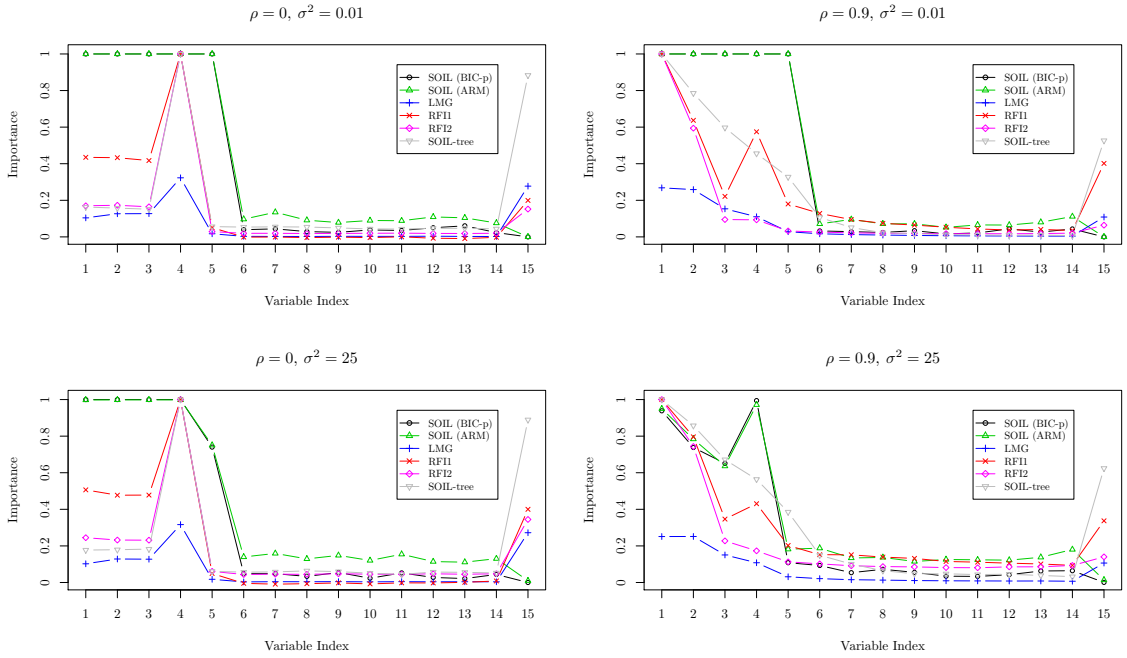


Figure 2.7: Simulation results for SOIL-tree on Example2

and $Q_m(T_k)$ is the deviance (node impurity if it is a classification tree) of the m-th 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 2.7 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.

2.6 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 2.3. We found that HT9 is the most important variable according to all methods. But different methods produce different second-most important variables.

	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**

Table 2.3: Importance measures of the variables in BGS data. The top two most important variables according to each measure are in bold.

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 et al., 2000; Rolling 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: $\mathbf{Y}_{new} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\sigma}N(0,1)$.

4. Compute the SOIL-ARM, SOIL-BIC-p, LMG, RFI1 and RFI2 importance measures using the new dataset $(\mathbf{X}, \mathbf{Y}_{new})$.

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.

The results are depicted in Figure 2.8. 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).

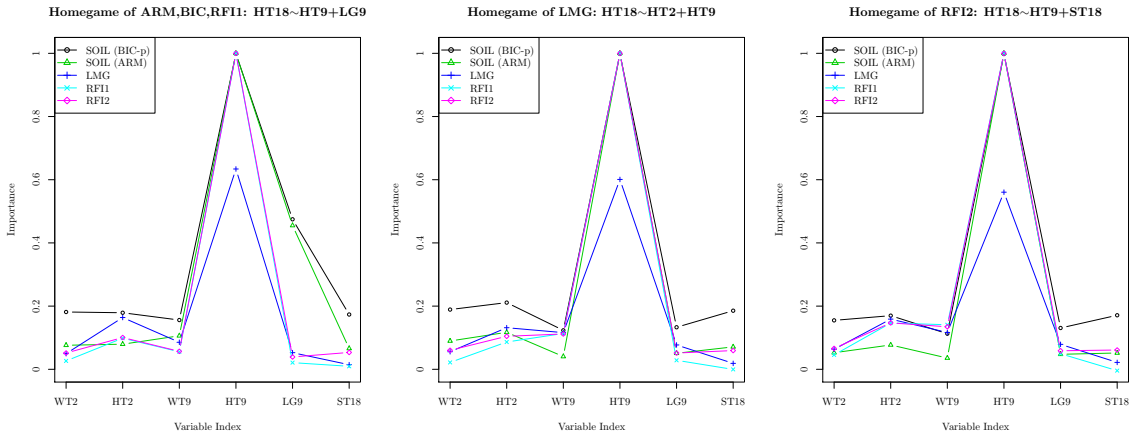


Figure 2.8: Results of cross-examination for BGS data.

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 2.4. 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 2.3. 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.

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

Table 2.4: Classical significance (p-value) analysis of the BGS data

Bardet data.

For a dataset with large p, we consider the Bardet dataset. It collects tissue samples from the eyes of 120 twelve-week-old male rats, which are the offspring of inter-crossed 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 et al., 2008a) is applied to the original probes, which gives us a dataset with 200 probes for each of 120 tissues. Specifically, 3000 out of the 31042 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 2.5. 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 2.5, we can see that different importance measures have very different results.

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

Table 2.5: Top ten genes for different variable importance measures for Bardet data.

Notice that X_{25141} is the most important variable according to Table 2.5. 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 2.5, the top 4 variables are selected for SOIL-BIC-p (SOIL-ARM), and the top 10 for RFI1 and RFI2 respectively.

In Figure 2.9, 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 3 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 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?

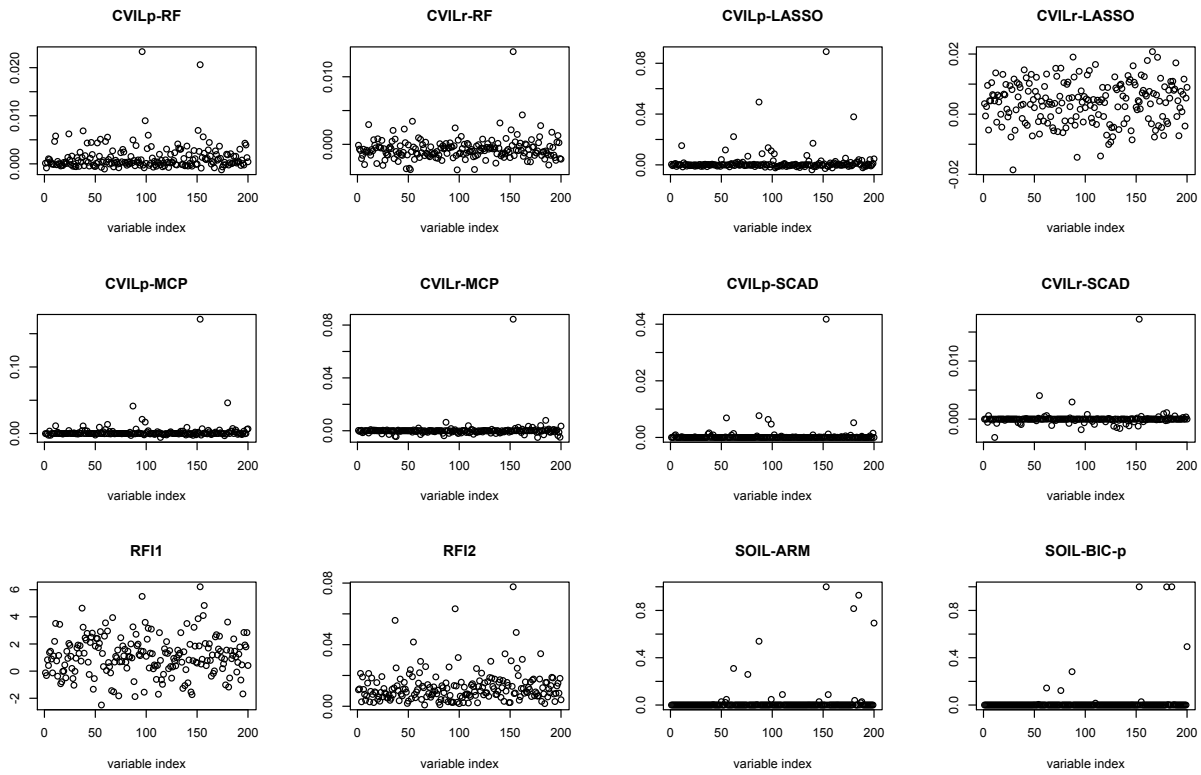


Figure 2.9: Simulation results for cross-examination

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 2.6 (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?

We present two SOIL importances also in Table 2.6. 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 2.6, 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

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)

Table 2.6: Top 5 variables for different variable importance measures of the Lung Cancer Data

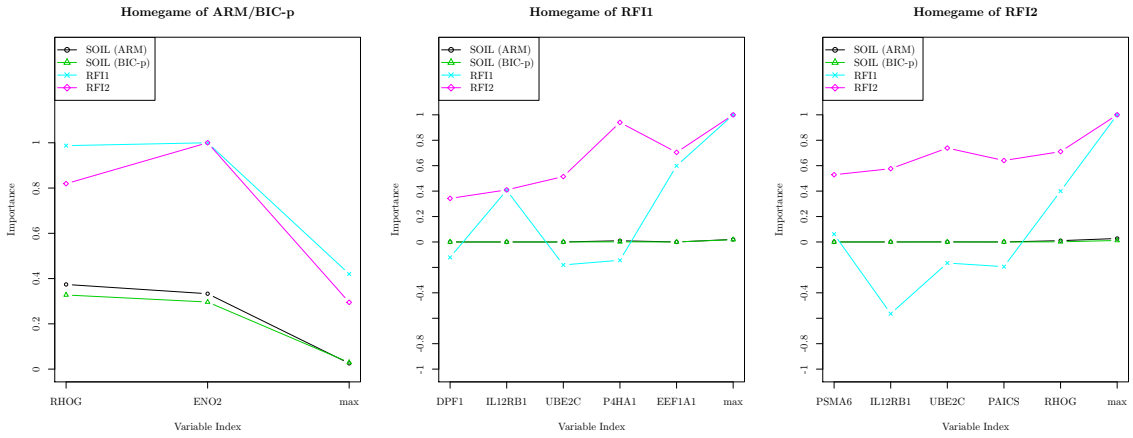


Figure 2.10: Results of cross-examination for Lung Cancer Data

vote proportion in random forest are utilized as the $\hat{p}$ for the home-game of SOIL and Random Forest respectively. Figure 2.10 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 non-parametric (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.

Chapter 3

Cross Validation Importance Learning

3.1 Introduction

Big data is ubiquitous nowadays, accompanied by numerous challenges. The interpretability of predictive variables/models is one challenge for many black-box methods, whose superior performances to traditional statistical methods in many applications are well recognized. The lack of interpretability hinder researchers/practitioners from applying these methods unreservedly.

Variable importance has been a popular methodology to demystify many currently prevalent black box methods. By understanding the marginal "importance" of each variable to the response, variable importance provides a non-parametric way of interpreting how a modeling procedure utilize each variable.

One major reason behind the adoption of variable importance is the generalization of machine learning and statistical methods. When a modeling method is overall unstable (or hard to be generalized), its performance on new-coming similar datasets or different types of datasets is not guaranteed by even an excellent performance of a specific dataset.

Various statistical techniques such as model selection (Tibshirani, 1996; Fan and Li, 2001; Zhang, 2010a) and machine learning (Cortes and Vapnik, 1995; Breiman,

2001; Rosenblatt, 1958) are applied to predict the output variable from an possibly enormous number of input variables.

Some methods such as model averaging use all the input variables, which is difficult for us to differentiate the effects of variables. Some methods have good predictive performance using only a small portion of the input variables. However, this sparsity adoption has given rise to two issues: first, the unselected variables are considered as making no contribution at all despite the fact that nearly all input variables have effect on the output variable to some extent, with many being negligible in prediction; second, an unstable procedure, given a slight change of the dataset, will lead to drastically different results of selected variables while maintaining a similar prediction performance. To improve the reliability of data analysis, it is desirable to have a robust and stable quantification of all the variables. Variable importance analysis is one way that gives the researchers/practitioners an overall understanding of the variables and thus helps determine which variables should be included in the model.

Many variable importance measures under linear regression have been proposed, such as regression coefficients, standardized regression coefficients, p-values, partial correlations. These measures fail to provide an evaluation of all the variables. Variance decomposition is another way of measuring the importance of variables in linear regression, including LMG (Lindeman et al., 1980), dominance analysis (Budescu, 1993), hierarchical partitioning (Chevan and Sutherland, 1991) and proportional marginal variance decomposition (PMVD) (Feldman et al., 1999). See Gromping (2007) for a more detailed review of variable importance measures based on variance decomposition. Ye et al. (2018) proposed sparsity oriented importance learning (SOIL), which incorporates a manageable set of candidate models with a sensitive weighting, and considers the sum of the weights of those candidate models that contains a certain variable as the importance of that variable.

Nonparametric variable importance measures in the literature lie in two major areas: random forest and causal inference. Breiman (2001) proposed random forest together with two types of variable importance measures. Many variants of the tree-

based variable importance measures are proposed later. For example, Strobl et al. (2008b) suggested a more reliable conditional permutation importance in random forest for correlated input variables. Sandri and Zuccolotto (2008) used the addition of pseudo-variables (Wu et al., 2007) to correct bias for the Gini variable importance measure in classification trees. Wang et al. (2010) used maximal conditional chi-square (MCC) to measure the conditional association between single nucleotide polymorphisms (SNPs) and the disease of interest. Chipman et al. (2010) and Bleich et al. (2014) proposed a **b**ayesian **a**dditive **r**egression **t**rees (BART) model and considered the proportion that each variable is used splitting rules of internal nodes within the trees as a variable selection approach. This variable inclusion proportion can also be viewed as a variable importance measure in BART model.

In contrast, in causal inference, variable importance is viewed as a real-valued parameter that is defined as the difference between the conditional mean of causal effect relative to the baseline. For example, $E(Y|A = a)-E(Y|A = 0)$ with covariates $X = (A, W)$ is defined as the marginal variable importance in Van der Laan (2006), where A is the variables of interest. In the framework of causal inference, such variable importance measure allow statistical inference including p-value and the confidence interval. For example, Van der Laan (2006) developed double robust estimators of several proposed marginal and adjusted variable importances. Hejazi et al. (2017) proposed a targeted variable importance and developed the corresponding estimators via targeted maximum likelihood estimation (TMLE) for datasets with small sample sizes. As pointed out in Williamson et al. (2017), these variable importance measures may be difficult to interpret in applications. They proposed a variable importance that can be interpreted as the increased variance of the outcome variable from the addition of the variable in the conditional mean function.

Variable importance measures are also proposed under other frameworks than the random forest and causal inference frameworks. Ribeiro et al. (2016) proposed a novel variable importance tool (LIME) to interpret and explain the predictions of classifiers at any fixed point. The idea behind LIME is that any function can be

locally approximated by linear functions. By solving a distance-weighted penalized regression at a local sample around the point of interest, LIME outputs the important covariates that explains the predictions provided by any uninterpretable classifier, helping people decide whether to trust the classifier for long-term purposes. Fisher et al. (2018) studied the variable importance to any model class, rather than a single model. The proposed model reliance (MR) investigates the expected loss of a model f on a switched sample, in which the covariates of interest in a sample is replaced by the same covariates from another independent sample with the same distribution. MR can be related to the conditional causal effects if the covariate is binary. The model class reliance (MCR) is then defined as the maximum and the minimum of the model reliance over a class of models. Through MCR, finite sample bounds and coverage properties are obtained for inference of the best model that minimize the expected prediction loss. For more references, we refer the readers to Olden et al. 2004; Huang et al. 2008b; Chambaz et al. 2012; Sapp et al. 2014. Many other fields have also used the variable importance to analyze problems, through different terminologies such as sensitivity analysis, screening methods, delta index, among others. For variable importance measures in areas other than statistics, see Wei et al. (2015) for a general review.

All the aforementioned variable importance measures are either difficult to interpret under many applications or associated tightly with specific methods. In addition, many machine learning methods do not usually provide a variable importance measure or insights of their predictions. This has motivated us to come up with Cross Validation Importance Learning (CVIL), which can be applied to any statistical model to provide interpretable variable importance measures of the variables in those "magical" machine learning methods, from the perspective of prediction. Based on the fixation of a variable at a constant value in the dataset or the deletion of a particular variable in the dataset, we propose two types of CVILs, CVIL_{p} and CVIL_{r}, by calculating the cross-validation averaged difference in prediction error (the sum of squared error) between the new dataset and the original dataset.

The contributions of this paper are three-fold. First, our method is general and can be applied to any modeling procedure. It enables one to interpret how important an input variable is considered by a specific modeling procedure in predicting the output variable. The proposed variable importance is interpreted as the proportion of improved prediction error after fixing or deleting a variable. Second, the two types of variable importance measures, CVIL_p and CVIL_r, evaluate the importance of each variable to the modeling procedure from the perspective of prediction accuracy (or position) and replaceability respectively. Under mild conditions, consistency results of both types of variable importances are established, together with the corresponding confidence intervals. Given any specific method, CVIL ranks the relative contribution of all input variables and provides us a way to interpret many predictive algorithms that lack interpretability.

We were aware of the work by Lei et al. (2018) during the write-up of this paper. Lei et al. (2018) proposed a model-free variable importance measure, named leave-one-covariate-out (LOCO), as a random parameter to measure the importance of a variable, under the conformal inference framework. The idea of LOCO is similar with our proposed *replaceability* variable importance, which is that deleting an important variable will decrease the prediction accuracy. In their follow-up paper (Rinaldo et al., 2016), the authors further investigate the theoretical properties of LOCO for post-selection inference. Though similar ideas, our *replaceability* importance and the LOCO importance differ in several aspects: 1) The LOCO variable importance is a random quantity conditional on the training set. As pointed out by the authors, for each splitting, the inference targets a different parameter due to its randomness. However, our *replaceability* importance measure is a fixed parameter that does not depend on the training set but focuses more on the general predictive importance of whether a variable can be replaced by other variables; 2) Due to the technical considerations in their proof, the authors redefine the LOCO parameter by adding a random noise that makes the inference conservative, as well as truncating the prediction values of the modeling procedure to make sure the LOCO is element-

wise bounded. Thus LOCO is different from our *replaceability* importance; 3) LOCO in finite sample inference investigates the local/conditional importance of a variable, which only requires a uniform boundedness assumption on the distribution family and can be estimated accurately. In contrast, our method focuses more on using the variable importance measures as a tool to assist the understanding of how any modeling methods employ the variables.

The remainder of this paper is as follows. In Section 3.2, we propose two types of variable importances, behind which the intuition is demonstrated by two theoretical examples. Section 3.3 presents the methodology of CVIL and its asymptotic properties. We also provide statistical inference of the CVIL importances in Section 3.4. Good performances of our proposed CVIL importances are illustrated by simulations in Section 3.5, and Section 3.6 presents the application of CVIL on one real data example. The proofs and some figures are included in the Appendix.

3.2 Variable Importance (VI)

We consider the data generating model

$$Y = f(\mathbf{X}) + \epsilon, \tag{3.1}$$

where $f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) : \mathbb{R}^p \mapsto \mathbb{R}$ is an unknown regression function, $\mathbf{X} = (X^1, ..., X^p)$ is a p-dimensional predictor and ϵ is the random noise (independent of $\mathbf{X}$) with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 < \infty$. Let $Z = (\mathbf{X}_i, Y_i)_{i=1}^n$ denote the dataset of n i.i.d. copies from the data generating model and denote the distribution of $\mathbf{X}$ as $P_{\mathbf{x}}$. Denote $||f(x)||_q = (\int |f(x)|^q dP_{\mathbf{x}})^{1/q}$ as the L_q norm for $q > 0$ with respect to the probability measure $P_{\mathbf{x}}$ and $||f||_\infty = \operatorname{ess\,sup} |f| = \inf\{c \geq 0 : |f(x)| \leq c \text{ almost surely}\}$ as the L_∞ norm.

The next two subsections introduce our proposals of two types of variable **i**mportances: the *position* variable importance and the *replaceability* variable importance.

3.2.1 The position variable importance measures

The idea is similar with the permutation variable importance measure in random forest Breiman (2001). Intuitively, given that every other variable stays unchanged, if a variable is important in predicting the response variable for a modeling procedure, breaking the connection between the variable and the response should increase the prediction error. There are many ways of undermining the information contained in a variable, such as permuting the columns of this variable in the training set (Breiman, 2001), deleting the column of the variable (Rinaldo et al., 2016). In terms of prediction,

We use the superscripts $\mathbf{X}^{(j)}$ to denote that the j-th covariate of the vector $\mathbf{X}$ is replaced with a constant c_j. Let δ be a modeling procedure that provides an estimator, denoted as $\hat{\delta}_n(\mathbf{x})$, of the mean regression function $f(\mathbf{x})$, where the subscript n indicates the number of observations used in the modeling procedure. Before stating the definition of the *position* variable importance, we need the following condition of the limiting behavior of the modeling procedure δ:

(A1) There exists a function $g_{\delta,n}(\mathbf{x})$ such that for any $j = 1, ..., p$, $||\hat{\delta}_n(\mathbf{X}^{(j)}) - g_{\delta,n}(\mathbf{X}^{(j)})||_2 \xrightarrow{p} 0$ as $n \to \infty$.

(A2) There exists a $g_\delta(\mathbf{x})$ such that $||g_{\delta,n}(\mathbf{X}^{(j)}) - g_\delta(\mathbf{X}^{(j)})||_2 \to 0$ as $n \to \infty$.

The convergence in probability is over the joint distribution of the $n$sample points that are used to fit the procedure δ. From the conditions, we emphasize here the limiting behavior of δ in high-dimensional cases, rather than its accurate prediction. Condition (A1) implies that the limiting function, $g_{\delta,n}(\mathbf{x})$, is allowed to depend on the sample size n. We still require the existence of a fixed function $g_\delta(\mathbf{x})$ as in condition (A2). Note that the function $\hat{\delta}_n$ contains the data points $\{\mathbf{X}_i\}_{i=1}^n$.

Definition 3 (*position* variable importance) Under the setting of (4.10), with conditions (A1)-(A2) hold, the *position* variable importance of the covariate X_j with

respect to the modeling procedure δ is defined as

$$\mathrm{VI}_{\mathrm{p}}(X_j; \delta, n) := \frac{\mathrm{E}(g_{\delta,n}(\mathbf{X}^{(j)}) - Y)^2}{\mathrm{E}(g_{\delta,n}(\mathbf{X}) - Y)^2} - 1. \qquad \square$$

Remark 4 VI_{p} focuses on the "position" of a variable in terms of the predictive performance. It computes the relatively increased prediction error after data perturbation of a variable on the test set, i.e., replacing the variable with a constant in the test set. The larger VI_{p} is, the more important a variable is in predicting the output variable with respect to the modeling procedure δ. It is a conditional variable importance since all the other variables stay unchanged. $\square$

Remark 5 Both types of variable importance measures are interpreted as the percent of increment of prediction error after changing a covariate. Our methods emphasize the importance of a variable in predicting the output variable rather than model identification. In practice, we can take the sample mean over the training set as an estimate of c_j. $\square$

Remark 6 The way of defining variable importance as a ratio instead of difference ensures that the VI_{p} is linear-transformation invariant. So the variable importances of one variable given by different modeling procedures are comparable. $\square$

3.2.2 The replaceability variable importance measures

We use the superscripts $\mathbf{X}^{(-j)}$ to denote that the j-th covariate of the vector $\mathbf{X}$ is deleted. Let $\hat{\delta}_n^{(-j)}(\mathbf{x}^{(-j)})$ be the estimator generated by the modeling procedure δ based on the variables $\mathbf{X}^{(-j)}$ of n observations, i.e., $\{\mathbf{X}_i^{(-j)}, Y_i\}_{i=1}^n$. The superscript $(-j)$ on $\hat{\delta}_n^{(-j)}$ is to emphasize its difference from the function $\hat{\delta}_n$, the domain of which is p-dimensional. We need the following conditions before stating the definition of *replaceability* variable importance:

(A3) For $j = 1, ..., p$, there exists a function $g_{\delta,n}^{(-j)}(\mathbf{x}^{(-j)})$ such that $||\hat{\delta}_n^{(-j)}(\mathbf{X}^{(-j)}) - g_{\delta,n}^{(-j)}(\mathbf{X}^{(-j)})||_2 \xrightarrow{p} 0$ as $n \to \infty$.

(A4) For $j = 1, ..., p$, there exists a function $g_{\delta}^{(-j)}(\mathbf{x}^{(-j)})$ such that $||g_{\delta,n}^{(-j)}(\mathbf{X}^{(-j)}) - g_{\delta}^{(-j)}(\mathbf{X}^{(-j)})||_2 \xrightarrow{p} 0$ as $n \to \infty$.

Definition 4 (*replaceability* variable importance) The *replaceability* variable importance ($\mathrm{VI_r}$) of a covariate X_j with respect to the modeling procedure δ is defined as

$$\mathrm{VI_r}(X_j; \delta, n) := \frac{\mathrm{E}(g_{\delta,n}^{(-j)}(\mathbf{X}^{(-j)}) - Y)^2}{\mathrm{E}(g_{\delta,n}(\mathbf{X}) - Y)^2} - 1. \qquad \square$$

Remark 7 In comparison to $\mathrm{VI_p}$, $\mathrm{VI_r}$ concentrates on the predictive performance when a variable is wiped out in the training data. It is a common issue in real life that many variables are highly correlated with each other. If modeling procedures (e.g., LASSO, Random Forest) only care about making a good prediction, then correlated variables might receive similar low *replaceability* variable importance values, since deleting one variable does not largely affect the prediction accuracy. From the perspective of replaceability, it is worth incorporating $\mathrm{VI_r}$ as another variable importance measure. Though $\mathrm{VI_r}$ is defined from a prediction perspective, it still provides an understanding of how different modeling procedures utilize highly correlated variables. The larger $\mathrm{VI_r}$ is, the more irreplaceable a variable is in predicting the output variable with respect to the modeling procedure δ. $\square$

3.2.3 Examples

Through the following theoretical examples, we illustrate the rationale behind the variable importances $\mathrm{VI_p}$ and $\mathrm{VI_r}$ for both parametric and non-parametric cases.

(1) Parametric case Let the data generating process be $Y = X_1 + X_2 + \epsilon_0$, where $\mathrm{E}X_1 = 0$, $\mathrm{Var}(X_1) < \infty$ and $\epsilon_0 \sim N(0, \sigma_0^2)$. Consider the following three cases:

(i) We have an additional variable $X_3 = X_1 + \epsilon_1$, where $\epsilon_1 \sim N(0, \sigma_1^2)$. Let δ be the modeling procedure that fits a linear regression of Y on the covariates (X_1, X_2, X_3).

(ii) We have an additional variable X_3 that is independent of X_1. Let δ be the modeling procedure that fits a linear regression of Y on the covariates (X_1, X_2, X_3).

(iii) We have an additional variable $X_3 = X_1 + X_2$. Let δ be the modeling procedure that chooses the model with the smallest BIC among linear regression models of Y on all possible subsets of (X_1, X_2, X_3).

Under proper conditions, we have that the estimated function by the linear regression δ converge to a function $g_\delta(\mathbf{x})$ as in Table 3.1. The theoretical importances of the variables X_1 and X_3 are presented in Table 3.2. In the *position* variable importance ($\mathrm{VI_p}$), the constant c_j is chosen accordingly so that the effect of the variable is the smallest among all possible c_j, i.e., $c_j = \arg\min_{c_j} \mathrm{VI_p}(X_j; \delta, c_j)$.

Case	$g_\delta(\mathbf{x})$	$g_\delta^{(-1)}(\mathbf{x}^{(-1)})$	$g_\delta^{(-3)}(\mathbf{x}^{(-3)})$
(i)	$x_1 + x_2$	$x_2 + bx_3$	$x_1 + x_2$
(ii)	$x_1 + x_2$	x_2	$x_1 + x_2$
(iii)	x_3	x_3	$x_1 + x_2$

Table 3.1: The limiting functions in the parametric case. Here the constant $b = \frac{\mathrm{Var}(X_1)}{\mathrm{Var}(X_1)+\sigma_1^2}$.

Case	$\mathrm{VI_p}(X_1)$	$\mathrm{VI_r}(X_1)$	$\mathrm{VI_p}(X_3)$	$\mathrm{VI_r}(X_3)$
(i)	$\frac{\mathrm{Var}(X_1)}{\sigma_0^2}$	$\frac{1}{\sigma_0^2}\frac{\sigma_1^2 \cdot \mathrm{Var}(X_1)}{\mathrm{Var}(X_1)+\sigma_1^2}$	0	0
(ii)	$\frac{\mathrm{Var}(X_1)}{\sigma_0^2}$	$\frac{\mathrm{Var}(X_1)}{\sigma_0^2}$	0	0
(iii)	0	0	$\frac{\mathrm{Var}(X_3)}{\sigma_0^2}$	0

Table 3.2: The variable importances of X_1 and X_3 in the parametric case.

Different patterns of the combination of VI_p and VI_r have different implications. For a variable, if $\text{VI}_\text{p} = \text{VI}_\text{r} > 0$, the role of this variable cannot be replaced by another variable for a static (without variable selection) modeling method δ, which can be seen in case (ii) for X_1. If $\text{VI}_\text{p} = \text{VI}_\text{r} = 0$, the variable can either be independent of any other variable or a confuser variable (correlated with a variable that is in the data generating process), which can be seen in case (i) and (ii) for X_3. In case (i), even if X_3 is correlated with X_1, it does not has a position in $g_\delta(\mathbf{x})$, which is consistent with its zero VI_p with respect to δ. If $\text{VI}_\text{p} > \text{VI}_\text{r} = 0$, the variable is replaceable by other variables (or may totally dependent with some other variables) but treated as irreplaceable by the modeling procedure δ, which can be seen in case (iii) for X_3. If $\text{VI}_\text{p} \neq \text{VI}_\text{r} > 0$ with VI_r small, the variable is important in predicting the response but is replaceable by another variable (without it, the predictive performance of δ may not change much), which can be seen in case (i) for X_1. In case (i), the VI_r importance of X_1 will be close to 0 when it is highly correlated to X_3 (σ_1^2 is small).

It is worth noticing that, for $\text{VI}_\text{r}(X_1)$ in case (i), $\frac{\sigma_1^2 \cdot \text{Var}(X_1)}{\text{Var}(X_1)+\sigma_1^2}$ is an increasing function of σ_1^2 and thus $\text{VI}_\text{r}(X_1)$ is upper bounded by $\frac{\text{Var}(X_1)}{\sigma_0^2}$ that happens to be the replaceability importance of X_1 in case (ii). This coincidence can be explained by the intuition that the replaceability importance of a variable reaches its maximum when no variable can replace it (i.e., $\sigma_1^2 \to \infty$ in our examples or in general all the variables are independent with the target variable).

(2) Nonparametric case Let the data generating process be $Y = f(X) + \epsilon_0 = \sum_{i=0}^{\infty} a_i \phi_i(X) + \epsilon_0$, where $X \sim \text{Uniform}(-1, 1)$ and $\epsilon_0 \sim N(0, \sigma_0^2)$. Let $P\phi_i(x)$ be the solution to $\min_t \text{E}(t - \phi_i(x))^2$, where t is a linear combination of the basis $\{\phi_j(x)\}_{j=0, j\neq i}^{\infty}$. Consider the following two cases:

(i) Non-orthogonal basis: $f(x) = e^x$ with $a_i = \frac{1}{i!}$ and $\phi_i(x) = x^i$.

(ii) Orthogonal basis: $f(x) = \frac{1}{\sqrt{2-2x}}$ with $a_i = 2^i$ and$\{\phi_i(x)\}_{i=0}^{\infty}$, where $\{\phi_i(x)\}_{i=0}^{\infty}$ are the Legendre polynomials.

For both (i) and (ii), let δ be the polynomial regressions of y on x, which selects the

order of the polynomials using AIC. Under proper conditions, we have the estimated function converges to $g_\delta(X) = \sum_{i=0}^{\infty} a_i \phi_i(x)$.

Case	$\mathrm{VI_p}(\phi_i(x))$	$\mathrm{VI_r}(\phi_i(x))$
(i)	$\frac{\mathrm{E}(\phi_i(x))^2}{\sigma_0^2 i! i!}$	$\frac{\mathrm{E}(P\phi_i(x)-\phi_i(x))^2}{\sigma_0^2 i! i!}$
(ii)	$\frac{2^{2i}}{\sigma_0^2}(\frac{1}{2i+1})$	$\frac{2^{2i}}{\sigma_0^2}\frac{1}{2i+1}$

Table 3.3: The variable importances of $\phi_i(x)$ in the nonparametric case. In VI_p, for the i-th base term $\phi_i(x)$, the constant c that replaces the variable is chosen as such that $\phi_i(c) = 0$, where $c$depends on i.

For any basic term $\phi_i(x)$ in the orthogonal basis case, when the constant c is properly selected such that $\phi_i(c) = 0$, the position importance and replaceability importance of $\phi_i(x)$ are the same (non-zero). In the non-orthogonal basis case, when a basic term, for example $\phi_{10}(x)$, is highly correlated with other basic terms, its $\mathrm{VI_r}$ importance is possibly close to 0 ($P\phi_i(x)$ is close to $\phi_i(x)$) but its $\mathrm{VI_p}$ importance is nonzero. Similar interpretation of the combination of $\mathrm{VI_p}$ and $\mathrm{VI_r}$ can be obtained as in the parametric case.

3.3 Cross-Validation Importance Learning (CVIL)

We use cross-validation to estimate the *position* and the *replaceability* variable importances. Cross-validation is a widely general method to evaluate and compare the predictive performances of modeling procedures on unknown data. In cross-validation, we randomly split the data into the training set for model fitting and the test set for model evaluation. However, predictive performance measured on only one split of the data is usually considered unstable. Different methods of cross-validations with multiple data splittings are thus proposed, which can be divided into two main categories: exhaustive cross-validation (e.g. leave-p-out CV; leave-one-out CV) and non-exhaustive cross-validation (e.g. k-fold CV; Monte Carlo CV; repeated learning-testing). The exhaustive cross-validation averages over all possible splittings while

non-exhaustive cross-validation averages over only a subset of all data splittings.

We introduce notations before stating our proposed cross-validation based estimates of the variable importances. Denote a random permutation of the observations as π_k, the first n_1 observations as the training set, the rest n_2 observations being the test set, i.e., $\mathbf{Z_1} = (\mathbf{X}_i, Y_i)_{i=1}^{n_1}$ and $\mathbf{Z_2} = (\mathbf{X}_i, Y_i)_{i=n_1+1}^{n}$. We denote $\mathbf{Z}_1^{(j)} = (\mathbf{X}_i^{(j)}, Y_i)_{i=1}^{n_1}$ and denote $\mathbf{Z}_1^{(-j)}$, $\mathbf{Z}_2^{(j)}$ and $\mathbf{Z}_2^{(j)}$ in a similar way. Let $\hat{\delta}_{n_1}(\mathbf{x})$ and $\hat{\delta}_{n_1}^{(-j)}(\mathbf{x}^{(-j)})$ denote an estimator of $f(\mathbf{x})$, generated by a modeling procedure δ on the training sets $\mathbf{Z_1}$ and $\mathbf{Z}_1^{(-j)}$ respectively. Define

$$\mathrm{CVIL_p}(X^j; \delta, n, \pi_k) := \frac{\sum_{i=n_1+1}^{n} (\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - Y_i)^2 / n_2}{\sum_{i=n_1+1}^{n} (\hat{\delta}_{n_1}(\mathbf{X}_i) - Y_i)^2 / n_2} - 1 \tag{3.2}$$

and

$$\mathrm{CVIL_r}(X^j; \delta, n, \pi_k) = \frac{\sum_{i=n_1+1}^{n} (\hat{\delta}_{n_1}^{(-j)}(\mathbf{X}_i^{(-j)}) - Y_i)^2 / n_2}{\sum_{i=n_1+1}^{n} (\hat{\delta}_{n_1}(\mathbf{X}_i) - Y_i)^2 / n_2} - 1 \tag{3.3}$$

be the **c**ross-**v**alidation based **i**mportance **l**earning (CVIL) of the variable importance $\mathrm{VI_p}$ and $\mathrm{VI_r}$ respectively over one data splitting π_k. Consider a collection of data splitting $\{\pi_k\}_{k=1}^{K}$ with the same splitting ratio, we use $\frac{1}{K}\sum_{k=1}^{K} \mathrm{CVIL_p}(X^j; \delta, n, \pi_k)$ and $\frac{1}{K}\sum_{k=1}^{K} \mathrm{CVIL_r}(X^j; \delta, n, \pi_k)$ for a more stable measure in practice.

3.3.1 Consistency

We establish the consistency of the two proposed CVIL variable importance measures in this subsection. The following conditions are required.

(A5) The functions f and g_δ are bounded almost surely, i.e., $||f||_\infty < \infty$ and $||g_\delta||_\infty < \infty$.

(A6) For $j = 1, ..., p$, the functions $\hat{\delta}_n$, $\hat{\delta}_n^{(-j)}$, $g_{\delta,n}$,$g_{\delta,n}^{(-j)}$ are uniformly bounded for n. For example, $\exists\ M_1 > 0$ such that $||\hat{\delta}_n||_\infty < M_1$, $\forall\ n \geq 1$.

Remark 8 The boundedness conditions (A5)-(A6) are commonly used in the regression function estimation literature. □

Theorem 3 Suppose the conditions (A1), (A2), (A5), (A6) hold and $n_2 \to \infty$, the cross-validation based position variable importance for X^j by the modeling procedure δ is consistent: □

$$\frac{1}{K}\sum_{k=1}^{K} \mathrm{CVIL_p}(X^j; \delta, n, \pi_k) - \mathrm{VI_p}(X_j; \delta, n) \xrightarrow{p} 0$$

as $n \to \infty$.

Proof 3.1

The proof is in the Appendix. □

Theorem 4 Suppose conditions (A3)-(A6) hold and $n_2 \to \infty$, the cross-validation based replaceability variable importance for X^j by the modeling procedure δ is consistent: □

$$\frac{1}{K}\sum_{k=1}^{K} \mathrm{CVIL_r}(X^j; \delta, n, \pi_k) - \mathrm{VI_r}(X_j; \delta, n) \xrightarrow{p} 0$$

as $n \to \infty$.

Proof 3.2

The proof is in the Appendix. □

Remark 9 The functions $g_{\delta,n}, g_{\delta,n}^{(-j)}$ do not need to have an explicit expression of x. The theorems imply that the two variable importance measures mimic the proportion of change in prediction error for the underlying modeling procedure δ after a perturbation to the dataset. The perturbation refers to fixation and deletion respectively in $\mathrm{CVIL_p}$ and $\mathrm{CVIL_r}$. If the modeling procedure itself is not predictive, CVIL-δ may fail to give an overall informative evaluation of the covariates. □

Remark 10 The above two variable importance measures can be applied to any modeling procedure. Additionally, it provides a way of interpreting some mysterious methods such as "black-box" machine learning algorithms since CVIL reflects how the modeling procedure itself evaluates the predictive power of each variable in the dataset. □

3.4 Statistical Inference

In this subsection, we establish the asymptotic normality of CVIL under appropriate conditions and provide its corresponding confidence intervals. Denote $\mu_g = \mathrm{E}_{\mathbf{X}}(g_{\delta,n}(\mathbf{X}) - Y)^2$, $\mu_\delta = \mathrm{E}_{\mathbf{X}|\mathbf{Z}_1}(\hat{\delta}_{n_1}(\mathbf{X}) - Y)^2$, $\sigma_g^2 = \mathrm{Var}_X(g_{\delta,n}(\mathbf{X}) - Y)^2$ and $\sigma_\delta^2 = \mathrm{Var}_{\mathbf{X}|\mathbf{Z}_1}(\hat{\delta}_{n_1}(\mathbf{X}) - Y)^2$. Then define $\mu_{g^{(j)}} = \mathrm{E}_{\mathbf{X}}(g_{\delta,n}(\mathbf{X}^{(j)}) - Y)^2$ and $\mu_{g^{(-j)}} = \mathrm{E}_{\mathbf{X}}(g_{\delta,n}^{(-j)}(\mathbf{X}^{(-j)}) - Y)^2$, where the superscript (j) and $(-j)$ indicate the difference in the function g. Denote $\sigma_{g^{(j)},g} = \sigma_{g,g^{(j)}} = \mathrm{Cov}((g_{\delta,n}(\mathbf{X}^{(j)}) - Y)^2, (g_{\delta,n}(\mathbf{X}) - Y)^2)$. We denote $\mu_{\delta^{(j)}}$, $\mu_{\delta^{(-j)}}$, $\sigma_{g^{(j)}}$, $\sigma_{g^{(-j)}}$, σ_δ, $\sigma_{\delta^{(j)}}$, $\sigma_{\delta^{(-j)}}$, $\sigma_{g^{(-j)},g}$, $\sigma_{\delta^{(j)},\delta}$ and $\sigma_{\delta^{(-j)},\delta}$ in a similar way. For the above notations, we omit n and n_1 in the subscripts for presentation convenience, e.g., we use μ_g instead of $\mu_{g,n}$ and μ_δ instead of μ_{δ,n_1}. The following conditions are needed. We also define $\mu_g^0 = \mathrm{E}_{\mathbf{X}}(g_\delta(\mathbf{X}) - Y)^2$, $(\sigma_g^0)^2 = \mathrm{Var}_{\mathbf{X}}(g_\delta(\mathbf{X}) - Y)^2$, and $\mu_{g^{(j)}}^0$, $\mu_{g^{(-j)}}^0$, $\sigma_{g^{(j)}}^0$, $\sigma_{g^{(-j)}}^0$, $\sigma_{g^{(j)},g}^0$, $\sigma_{g^{(-j)},g}^0$in a similar way, where the superscript 0 emphasizes that the constant does not depend on n.

(B0) $n_2 \to \infty$.

(B1) $\sqrt{n_2} \cdot ||\hat{\delta}_{n_1}(\mathbf{x}) - g_{\delta,n_1}(\mathbf{x})||_1 \xrightarrow{p} 0$ as $n \to \infty$.

(B2) $\sqrt{n_2} \cdot ||\hat{\delta}_{n_1}(\mathbf{x}^{(j)}) - g_{\delta,n_1}^{(j)}(\mathbf{x}^{(j)})||_1 \xrightarrow{p} 0$ as $n \to \infty$.

(B3) $\sqrt{n_2} \cdot ||\hat{\delta}_{n_1}^{(-j)}(\mathbf{x}^{(-j)}) - g_{\delta,n_1}^{(-j)}(\mathbf{x}^{(-j)})||_1 \xrightarrow{p} 0$ as $n \to \infty$.

(B4) $\frac{1}{\sqrt{n_2}} E \left\| \begin{pmatrix} \sigma_{\delta^{(j)}}^2 & \sigma_{\delta,\delta^{(j)}} \\ \sigma_{\delta,\delta^{(j)}} & \sigma_\delta^2 \end{pmatrix}^{-\frac{1}{2}} \begin{pmatrix} (\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - Y_i)^2 - \mu_{\delta^{(j)}} \\ (\hat{\delta}_{n_1}(\mathbf{X}_i) - Y_i)^2 - \mu_\delta \end{pmatrix} \right\|^3 \to 0$ as $n \to \infty$, where $||\cdot||$ is the Euclidean norm.

(B5) $\frac{1}{\sqrt{n_2}}E\left\|\begin{pmatrix}\sigma^2_{\delta^{(-j)}} & \sigma_{\delta,\delta^{(-j)}}\\ \sigma_{\delta,\delta^{(-j)}} & \sigma^2_{\delta}\end{pmatrix}^{-\frac{1}{2}}\begin{pmatrix}(\hat{\delta}_{n_1}(\mathbf{X}_i^{(-j)})-Y_i)^2-\mu_{\delta^{(-j)}}\\ (\hat{\delta}_{n_1}(\mathbf{X}_i)-Y_i)^2-\mu_{\delta}\end{pmatrix}\right\|^3 \to 0$ as $n \to \infty$, where $||\cdot||$ is the Euclidean norm.

Remark 11 The above conditions put constraints on the splitting ratio n_2/n_1, based on the specific convergence rate of the modeling procedure. For example, the $||\hat{\delta}_{n_1}(\mathbf{x})-g_{n_1,\delta}(\mathbf{x})||_2^2$ is of order $1/n_1$ for parametric methods δ, thus the requirement of the splitting ratio is that $n_2/n_1 \to 0$. In practice, half-half splitting or 60/40 splitting usually has a good performance to our experience. □

Remark 12 Conditions (B4) and (B5) require that the standardized version of the prediction error $(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)})-Y_i)^2$, $(\hat{\delta}^{(-j)}_{n_1}(\mathbf{X}_i^{(-j)})-Y_i)^2$ and $(\hat{\delta}_{n_1}(\mathbf{X}_i)-Y_i)^2$ have third moments of order $O(\sqrt{n_2})$. These errors need not necessarily to be uniformly bounded. The conditions also require that $\text{corr}_{\mathbf{X}|\mathbf{Z}_1}((g_{n_1,\delta}(\mathbf{X}^{(j)})-Y)^2,(g_{n_1,\delta}(\mathbf{X})-Y)^2) \neq 1$ and $\text{corr}_{\mathbf{X}|\mathbf{Z}_1}((g^{(-j)}_{n_1,\delta}(\mathbf{X}^{(-j)})-Y)^2,(g_{n_1,\delta}(\mathbf{X})-Y)^2) \neq 1$. For the variables that are not used by a modeling procedure δ at all, conditions (B4) and (B5) no longer hold since $\sigma^2_{\delta}=\sigma^2_{\delta^{(j)}}=\upsilon_{\delta,\delta^{(j)}}$ and $\sigma^2_{\delta}=\sigma^2_{\delta^{(-j)}}=\upsilon_{\delta,\delta^{(-j)}}$. For these variables, the CVIL_p and CVIL_r importances are equivalent to 0. The results in this subsection focus on variables that are used by the modeling procedure δ. □

Theorem 5 (Asymptotic Normality of CVIL_p) Assume conditions (A2), (B0), (B1), (B2) and (B4) hold. Define $\text{VI}_\text{p}(X_j;\delta)=\frac{\text{E}(g_\delta(\mathbf{X}^{(j)})-Y)^2}{\text{E}(g_\delta(\mathbf{X})-Y)^2}-1$. We have

$$\sqrt{n_2}(\text{CVIL}_\text{p}(X^j;\delta,\pi_k)-\text{VI}_\text{p}(X_j;\delta)) \xrightarrow{d} N\left(0,\left(\frac{\sigma^0_{g^{(j)}}}{\mu^0_g}\right)^2+\left(\frac{\mu^0_{g^{(j)}}\sigma^0_g}{(\mu^0_g)^2}\right)^2-2\frac{\mu^0_{g^{(j)}}\sigma^0_{g,g^{(j)}}}{(\mu^0_g)^3}\right)$$

as $n \to \infty$. □

Proof 3.3

The proof is in the Appendix. □

We have similar results for CVIL_r as follows.

Theorem 6 (Asymptotic Normality of CVIL_r) Assume conditions (B0), (B1), (B3) and (B5) hold. Define $\text{VI}_\text{r}(X_j;\delta) = \frac{\text{E}(g_\delta^{(-j)}(\mathbf{X}^{(-j)})-Y)^2}{\text{E}(g_\delta(\mathbf{X})-Y)^2} - 1$. We have

$$\sqrt{n_2}(\text{CVIL}_\text{r}(X^j;\delta,\pi_k) - \text{VI}_\text{r}(X_j;\delta)) \xrightarrow{d} N\left(0, \left(\frac{\sigma^0_{g^{(-j)}}}{\mu^0_g}\right)^2 + \left(\frac{\mu^0_{g^{(-j)}}\sigma^0_g}{(\mu^0_g)^2}\right)^2 - 2\frac{\mu^0_{g^{(-j)}}\sigma^0_{g,g^{(-j)}}}{(\mu^0_g)^3}\right)$$

as $n \to \infty$. □

Proof 3.4

The proof is in the Appendix. □

One natural estimate of the standard deviation of the CVIL_p is to plug in the sample mean, sample variance and sample covariance. For instance, the "sample" mean estimate of μ^0_g based on the test sample is , i.e. $\hat{\mu}^0_g = \frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - Y_i)^2$. The other estimates are obtained in a similar way. Based on the asymptotic normality of the CVIL, we can build the corresponding confidence interval as follows. we can construct a $1-\alpha$ confidence interval for the VI_p importance of X^j as following:

$$\text{CVIL}_\text{p}(X^j;\delta,\pi_k) \pm z_{\alpha/2}\sqrt{\frac{1}{n_2}\left(\left(\frac{\hat{\sigma}^0_{g^{(-j)}}}{\hat{\mu}^0_g}\right)^2 + \left(\frac{\hat{\mu}^0_{g^{(-j)}}\hat{\sigma}^0_g}{(\hat{\mu}^0_g)^2}\right)^2 - 2\frac{\hat{\mu}^0_{g^{(-j)}}\hat{\sigma}^0_{g,g^{(-j)}}}{(\hat{\mu}^0_g)^3}\right)},$$

where $z_{\alpha/2}$ is the lower $\alpha/2$ quantile of the standard normal distribution. The $1-\alpha$ confidence interval for the VI_r importance of X^j can be constructed in a similar way.

Remark 13 This result only applies to CVIL based on one data splitting. To build a confidence interval for CVIL with multiple splittings, we can use the mean of the plug-in estimators over multiple splittings. For example, the point estimate in the confidence interval will be $\sum_{k=1}^{K}\text{CVIL}_\text{p}(X^j;\delta,\pi_k)$ and the estimate of u^0_g will be $\hat{\mu}^{0\prime}_g = \frac{1}{n_2}\sum_{k=1}^{K}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)};\pi_k) - Y_i)^2$.

3.5 Numerical Studies

In this section, we demonstrate the performance of CVIL_p and CVIL_r from various simulation settings. Modeling methods such as generalized additive model (GAM), neural network (NN), LASSO, linear regression with full model (LR) and stepwise linear regression (LRs) are evaluated. Another two methods (random forest (RF) and LMG), known for providing a variable importance measure within the methods, are also considered. Here LMG is a relative importance measure from linear regression which averages over all possible orderings of the variables when fitting a linear model. Random forest provides two types of variable importance measures, denoted as RFI1 and RFI2 in this paper. For each variable, RFI1 is the normalized difference between prediction errors based on pre-permuted and post-permuted out-of-bag data; RFI2 evaluates the total decrease in node impurity (Gini index for classification and residual sum of squares for regression) over all splits of the variable and all trees.

3.5.1 Simulations settings

We take into consideration various aspects of model settings, such as dimensions, data generating models and variable correlations. The design matrix $X_{n\times p}$ is generated from a zero-mean multivariate normal distribution, where the (i, j) entry of the covariance matrix is $\rho^{|i-j|}$, with $\rho = 0$ or 0.9. The noise variable ϵ is generated from a univariate normal distribution $N(0, 0.01)$. The model setups of the simulation examples are described in Table 3.4. Basically, the first three examples (Examples 1.1 to 4.3) are linear cases and the other three are generalized linear/additive models.

For low-dimensional linear cases, we compare the performance of CVIL_p/CVIL_r-δ ($\delta\in$\{RF,LR,LRs,LASSO\}) and three variable importance measures LMG, RFI1, RFI2. For the high-dimensional linear case (Example 1.2), only CVIL_p-δ ($\delta \in$ \{RF,LASSO\}) and RFI1/RFI2 are compared in the high-dimensional linear case where CVIL_r is time-consuming and LMG is computationally infeasible. In non-linear cases, we include CVIL_p/CVIL_r-δ ($\delta\in$\{GAM,RF,NN\}) and RFI1/RFI2.

For each specific simulation setting, five times of two-fold cross-validation are conducted in each repetition and we repeat the simulation 100 times. The averaged variable importance measures are then computed.

Example	n	p	model
Linear cases			
1	150	1000	$y = x_1 + x_2 + x_3 + \frac{1}{2}x_4 + \frac{1}{2}x_5 - \frac{1}{2}x_6 - \frac{1}{2}x_7 + \epsilon$
2	150	10+1	$y = x_1 + x_2 - \frac{1}{2}x_3 - \frac{1}{8}x_4 + \epsilon,\ x_{11} = x_1 + \epsilon_0,\ \epsilon_0 \sim N(0, 10e-6)$
3	150	10	$y = c(x_1 + x_2 + ... + x_{10}) + \epsilon,\ c = 1, 10, 0$
Nonlinear cases			
4	150	10	$y = \exp(x_1 + x_2 - x_3 + \frac{3}{4}x_4 - \frac{3}{4}x_5) + \epsilon$
5	150	10	$y = \exp(x_1) + \exp(x_2) + \sin(x_3) + x_4 + \frac{1}{2}x_5 - \frac{1}{4}x_6 + \frac{1}{6}x_7 + \epsilon$

Table 3.4: Model setups for simulation study. In this table, the dimension $p = 10 + 1$ indicates that we have 11 variables in total, among which 10 variables are generated from a 10-dimensional multivariate normal distribution and one variable is specifically generated (such as binomial distribution, linear combination of the first 10 variables). Examples A1.1, A3 and A6 are deferred to the Appendix.

3.5.2 Performance of the CVIL based importance measures

The results of the simulations are displayed in Figures (3.1- B.5). When plotting the random forest variable importances, we standardize RFI1 and RFI2 by dividing the maximum value of variable importances among p variables so that the maximum value for RFI1 and RFI2 is always 1 in the figures. For simplicity, we call the variables with non-zero coefficients "true" variables. We summarize the performance of the CVIL and other methods in the following aspects.

Example 1

Replaceability may help the discovery of causal relationship In Example 1, we want to investigate the performance of CVIL in the case where there exists a

replaceable true variable. In this example, X_{11} is not in the model but predictive due to its high correlation with the true variable X_1. From the CVIL$_{\text{p}}$-δ and CVIL$_{\text{r}}$-δ importances with δ being RF, LR, LRs or LASSO, we can make the conclusion that $\{X_1, X_2, X_3, X_{11}\}$ are predictive, among which X_{10} and X_1 are replaceable. Then we may consider to further investigate the relationship between these two replaceable variables if we are interested in finding out who is in the true model (or who has the causal relationship with the response rather than just correlation). But if we are only provided with LMG, RFI1 or RFI2, no conclusions of causal relationship can be made any further for X_{10} and X_1.

Example 2

Joint importance and marginal importance In Example 2 when $c = 1$, none of the confidence interval of CVIL$_{\text{p}}$-δ and CVIL$_{\text{r}}$-δ ($\delta \in \{$RF, LR, LRs, LASSO$\}$) variable importances contain 0, meaning all the variable are utilized by the modeling procedure δ and are not replaceable. However, for a given variable, say X_1, we have CVIL$_{\text{p}}$-$\delta(X_1)$ > CVIL$_{\text{r}}$-$\delta(X_1)$ for $\delta =$ RF, while CVIL$_{\text{p}}$-$\delta(X_1)$ < CVIL$_{\text{r}}$-$\delta(X_1)$ for $\delta =$ LR, LRs or LASSO. This is due to the difference of the modeling procedure. Note that the CVIL variable importances in our paper are from a marginal importance perspective. For example, in linear regression, when all the variables are independent with each other, deleting it (not using the information of the variable in both training and testing) will lead to larger prediction error than replacing it with a constant (using the information for training but not testing). However, RF is a modeling procedure that divide data into subgroups (terminal nodes), and replacing one variable with a constant actually destroys the joint relationship of the variables, thus the predictive performance is worse than that of deleting a variable.

Relativeness and absoluteness These three sub-examples ($c = 1, 10, 0$) in Example 2 are designed to interpret the scale of CVIL, as well as the relative-

ness/absoluteness of variable importance measure. Comparing sub-examples in the low correlation setting ($\rho = 0$), (Figures 3.2, 3.3 and 3.4), whether a set of equally-important (same values of coefficients) variables are absolutely important or relatively important is reflected by both the scale of CVIL and the modeling procedure δ. We notice that the scale of CVIL_p/CVIL_r and RFI2 increases dramatically when the coefficient β changes from zero (Figure 3.4) to non-zero (Figures 3.2 and 3.3). An easy interpretation of CVIL_p is the increased ratio of the prediction error when replacing the variable with the sample mean (note the scales of CVIL are different between $c = 1$ and $c = 10$). But for RFI, it is hard to interpret the scale since it only calculates the difference between pre-permuted and post-permuted data. Booth RFI1/RFI2 and CVIL_p/CVIL_r-RF imply that random forest cannot differentiate between $c = 0$ and $c \neq 0$. It can be concluded that random forest importance measures the relativeness rather than the absoluteness of the variable importance, but these two examples are totally different in practice. A relative variable importance measure might lead to misleading scientific findings.

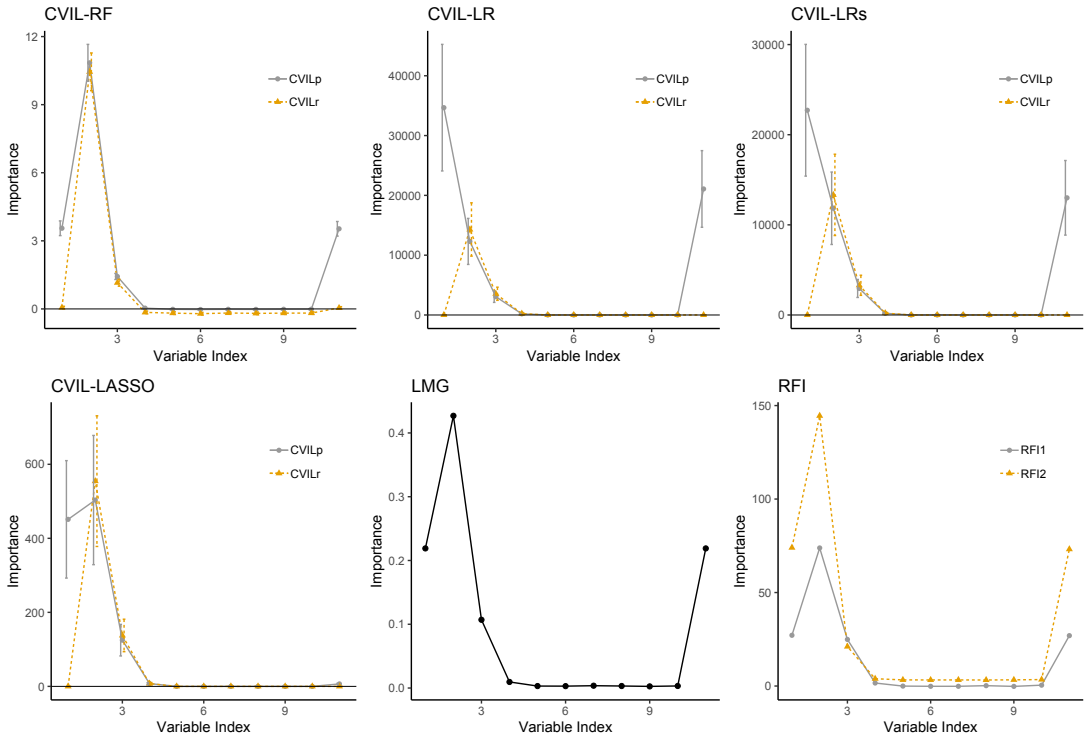
Improvement of CVIL-RF over RFI Comparing the performances of CVIL_p/CVIL_r-RF and RFI1/RFI2, it is easier to make conclusions of the "importance" of variables with the confidence intervals provided by CVIL. When $c = 1$ or $c = 10$, all the variables are relatively equally important based on CVIL or RF. With the confidence intervals provided by CVIL, CVIL improves the RFI by providing more details of whether the value is significantly different than 0, thus provide researchers/practitioners a clear cutoff (or equivalently reference like p-values) when employing the CVIL variable importance measures.

Example 3

Applicability to non-linear models Example 3 is designed for data generated from a generalized additive model and a generalized multiplicative model. By all the

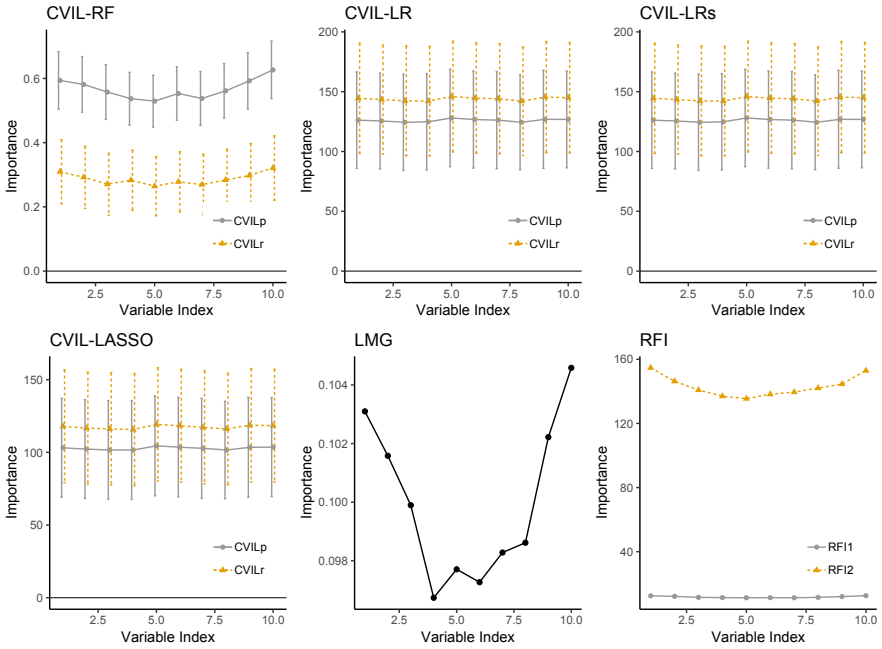
methods, only true variables with large coefficients ((x_1, x_2, x_3, x_4) for Example 3) are assigned large values of importance.

Negativity of the confidence interval In Example 3 (Figure 3.5), an interesting phenomena is that the confidence intervals of $\text{CVIL}_\text{p}/\text{CVIL}_\text{r}$-GAM for $\{X_{5,}..., X_{10}\}$ are all negative (all numbers in the interval are negative). We mark this as a characteristic of our method. When there is not enough sample size, fitting an extra unimportant (not in the model or has small predictive power) variable may even increase the prediction error. So we increase the sample size of Example 5 from $n = 150$ to $n = 300, 500$ and the confidence intervals begin to contain 0. The negativity of the confidence interval actually implies the "no/negative predictive power" of the variables under the current sample size, suggesting us to throw away these "definitely unimportant" variables.



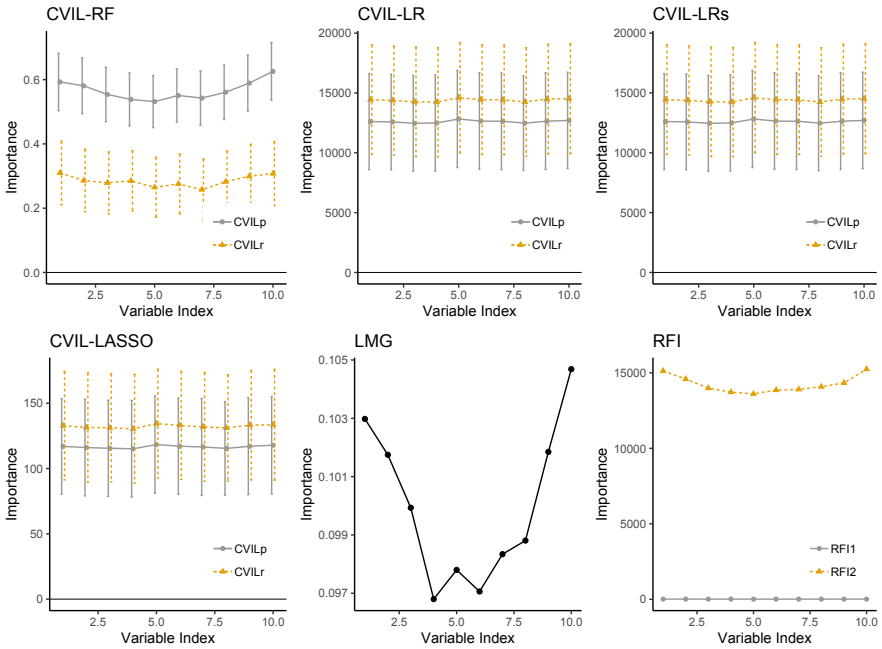
[H]

Figure 3.1: Example 1, $\rho = 0$. This example is to demonstrate the *replaceability* variable importance in terms of model selection/estimation.



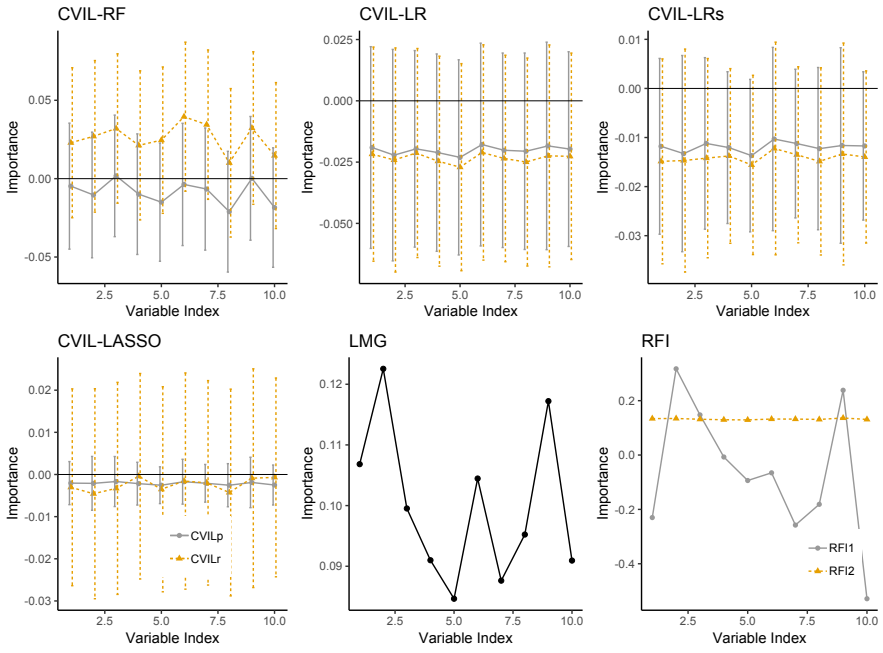
[H]

Figure 3.2: Example 2, $c = 1$, $\rho = 0$. This example is to demonstrate the absoluteness and relativeness of variable importance measures.



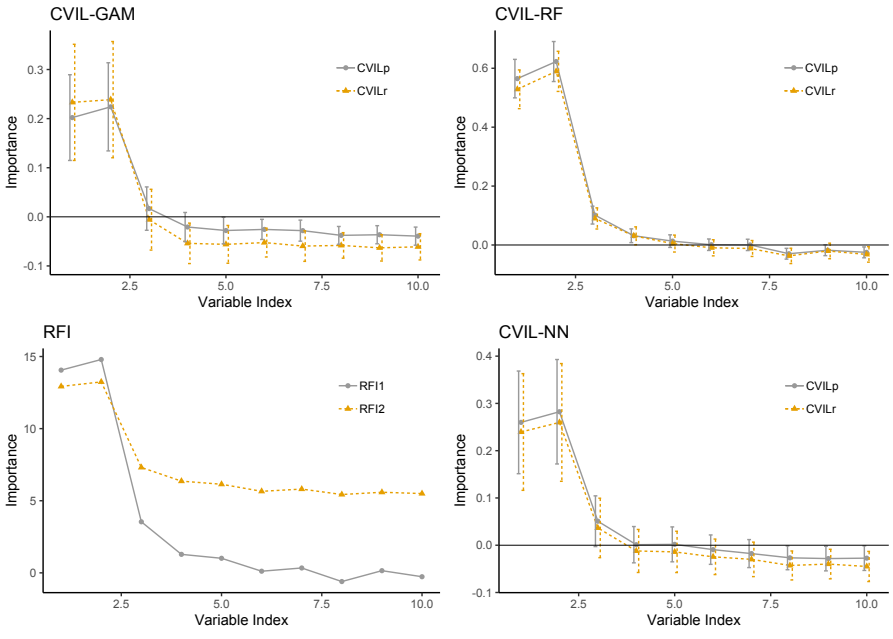
[H]

Figure 3.3: Example 2, $c = 10$, $\rho = 0$



[H]

Figure 3.4: Example 2, $c = 0$, $\rho = 0$



[H]

Figure 3.5: Example 3, $\rho = 0$

3.6 Real Data Examples

We investigate the performance of CVIL_p and CVIL_r of some model procedures δ on four real data applications. In the real data analysis, we obtain the averaged values of $\text{CVIL}_\text{p}/\text{CVIL}_\text{r}$ using two-fold cross-validation based on 100 repetitions. Direct variable importance measures such as RFI1/RFI2 and SOIL-ARM/BIC-p (Ye et al., 2018) are conducted for comparison.

3.6.1 Prostate cancer data

We consider the prostate cancer study in Stamey et al. (1989); Friedman et al. (2001), which investigates the correlation between the **p**rostate **s**pecific **a**ntigen (*psa*) level and a list of other 8 medical measurements in 102 patients before and after receiving a radical prostatectomy. The postoperative *psa* values are only available in 97 patients, so we use the dataset (available within the R package *lasso*2) used in Friedman et al. (2001), with 97 patients and 9 variables. The descriptions of the response *lpsa* and 8 predictors are presented in Table 3.5.

Variables	Description
X_1: *lcavol*	log of cancer volume
X_2: *lweight*	log of prostate weight
X_3: *age*	patient age
X_4: *lbph*	log of benign prostatic hyperplasia amount
X_5: *svi*	seminal vesicle invasion
X_6: *lcp*	log of capsular penetration
X_7: *gleason*	gleason score
X_8: *pgg45*	percentage gleason score
Y : *lpsa*	log of prostate specific antigen (*psa*) levels

Table 3.5: Variable description of prostate cancer data

In comparison to the analysis in Friedman et al. (2001), we follow their practice of treating all the variables as continuous. The residual analysis of the linear regression of the response on all the predictors suggests that the linear model is a good fit. In this example, to conduct the variable importance analysis, we consider the modeling procedure δ to be one of $\{\text{RF}, \text{LR}, \text{LRs}\}$. Additionally, we include four direct variable importance measures: RFI1, RFI2, SOIL-BIC-p, SOIL-ARM, the results of which are presented in Figure 3.6. Overall, CVIL_p and CVIL_r (with 95% confidence interval) have similar performances over different modeling procedures δ. All the CVIL methods agree on that *lcavol* is the only important variable in terms of prediction, indicating the high correlation between variables. For the performances of CVIL_p, particularly, the procedure "replacing variables with the training sample mean" when calculate the position importance is not effective enough due to the low variance of the variables *lweight* and *svi* in the dataset, hence their insignificance of position importance. An alternative approach could be to set a grid of constants and pick the one that achieves the largest position importance through cross validation. In the analysis of Friedman et al. (2001), four variables *lcavol*, *lweight*, *svi* and *lbph* are considered significant with Z-scores 5.37, 2.75, 2.47 and 2.06 respectively (Z-scores with its absolute value larger than 2.002 are considered significant). Their analysis is based on a training set with size 67 in one data split. So we fit a linear regression directly on the whole dataset and the results of Z-score show the significant variables are *lcavol*, *lweight* and *svi* under 0.05 significance level, which are the top 3 variables in each variable importance measure in Figure 3.6.

To demonstrate that CVIL can be used to find potential predictive interaction terms, we did a guided-simulation as following. First we generate a new response based on the following linear model with two-way interactions:

$$Y_{new} = 0.1X_1 + 0.1X_4 + 0.1X_6 + X_1X_6 + X_4X_6 + \epsilon,\ \epsilon \sim N(0, 0.01).$$

Then we obtain the variable importance measures, presented in Table 3.6, for each variable based on this new dataset $\{Y_{new}, X\}$. From Table 3.6, it can be observed that

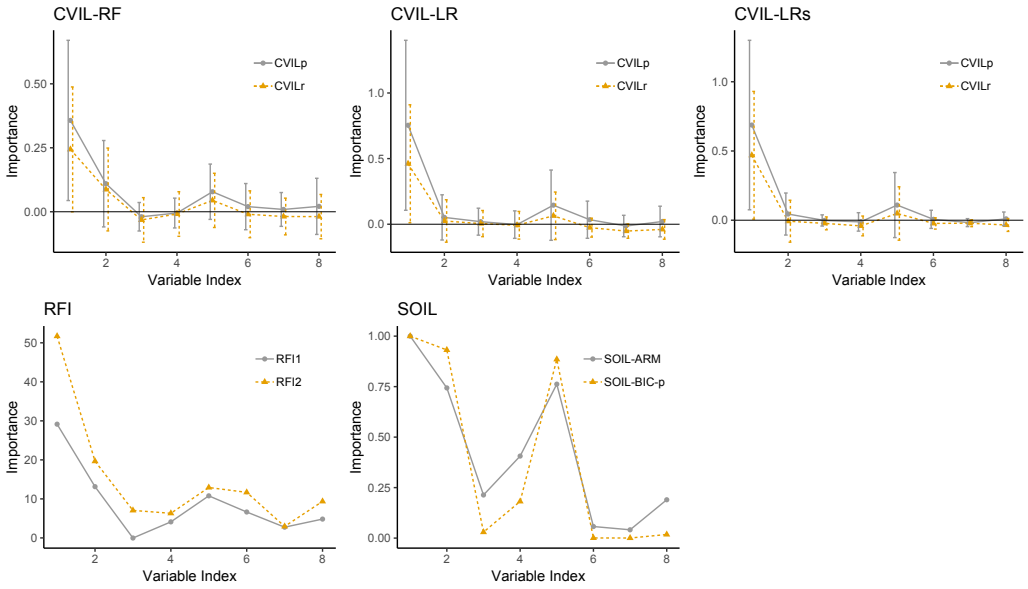


Figure 3.6: Importance measures of the prostate cancer data

it is hard for linear regression or CVIL-LR (with the response regressed on all the main effects of the variables) to detect all the three true variables from the perspective of prediction, especially when the coefficients of the main effects are small. It is not surprising to see that X_6 is the only important variable suggested by CVIL_p/CVIL_r-LR, since X_6 appears in both interaction effects X_1X_6, X_4X_6. In comparison, RFI1/RFI2 (if we pick the first three variables) and CVIL_p/CVIL_r-RF can identify the correct main effects even these variable importances are designed in terms of prediction power. One reason of the success of RF is the tree structure naturally includes interactions between variables. It is also worth mentioning that, unlike RFI1/RFI2 (we need to decide the number of variables to be selected), CVIL provides statistical inference tool such as the confidence interval to exclude the unimportant main effects. Thus, by comparing the different performances of LR and RF (and the corresponding variable importances), researchers/practitioners can move further to investigate interactions effects.

	LR (p-value)	$CVIL_p$-LR	$CVIL_r$-LR	RFI1	RFI2	$CVIL_p$-RF	$CVIL_r$-RF
X_1	**8.33e-05**	0.32	0.16	31.17	97.96	**0.33**	0.18
X_2	0.39	-0.01	-0.03	2.17	24.19	0.02	-0.04
X_3	0.60	-0.04	-0.04	-0.04	13.03	-0.01	-0.06
X_4	0.10	0.02	0.00	49.27	251.49	**0.71**	**0.58**
X_5	0.59	-0.05	-0.04	2.45	1.77	-0.01	-0.01
X_6	**1.84e-14**	**2.07**	**0.87**	61.19	856.70	**4.02**	**1.49**
X_7	0.28	0.02	-0.01	-1.03	0.57	0.00	0.01
X_8	0.12	0.03	-0.02	6.25	14.63	0.05	-0.06

Table 3.6: Variable importance measure for the guided simulation of the prostate data. The highlighted values are either p-values that are less than 0.05 or CVIL importances whose 95% CI doesn't contain 0.

Chapter 4

High-dimensional Adaptive Minimax Sparse Estimation with Interactions

4.1 Introduction

High-dimensional data are increasingly prevalent in various areas such as bioinformatics, astronomy, climate science and social science. When the number of variables p is larger than the sample size n in the linear regression setting, statistical estimation of the regression function often requires some crucial conditions. One common condition is the sparsity of the data generating model, under which only a small portion of the variables are important to affect the response variable. Under this condition, both sparse estimation of high-dimensional linear regression functions and variable selection have been well studied with fruitful theoretical understandings in the recent decade. Minimax estimation of the regression function with main effects only are well investigated under l_q-sparsity constraints with $0 \leq q \leq 1$ Candes and Tao (2007); Bunea et al. (2007); Zhang and Huang (2008); Van De Geer and Bühlmann (2009); Bickel et al. (2009); Zhang (2010b); Knight and Fu (2000); Raskutti et al. (2011); Rigollet and Tsybakov (2011); Wang et al. (2014); model selection consistency results are also obtained for various model selection procedures Fan and Li (2001); Zhao and Yu (2006); Zhang and Huang (2008); Lv and Fan (2009).

However, models with only main effects are often not adequate to fully capture the nature of the data. Interaction terms may be necessary to not only improve the prediction performance but also enhance the understanding of the relationships among the variables, especially in areas such as social networks, medicine, and genetics, where interaction effects between the covariates are of enormous interest. Hierarchical constraints are often imposed to describe the underlying structure of models with interaction effects, such as the marginality principle Nelder (1977), the effect heredity principle Hamada and Wu (1992) and the "well-formulated models" Peixoto (1987). We follow a popular naming convention of heredity conditions as adopted in Chipman (1996): strong heredity and weak heredity. Strong heredity assumes that if an interaction term is in the model, then both of its corresponding main effects should also be included, while weak heredity only requires that at least one of its main effects should be included. In practice, it is possible that, compared to the interaction terms, some main effects are so small that including them in modeling may not be beneficial from the perspective of estimation variability. Thus, in this work we take into consideration the additional case where no heredity condition is imposed at all, also for the purpose of theoretical comparison with the other two heredity conditions.

Many approaches are proposed for interaction selection, most of which can be categorized into two types: *joint selection* and *stage-wise selection*. The joint selection approach selects the main and interaction terms simultaneously by searching over all possible models with interactions. A typical way of joint selection is to use regularization methods with specially designed penalty terms. For example, Yuan et al. Yuan et al. (2009) introduced a family of shrinkage estimators, which incorporate the hierarchical structures through linear equality constraints on the coefficients and possess both selection consistency and root-n estimation consistency under fixed p. Choi et al. Choi et al. (2010) re-parameterized the regression model with interactions and applied an adaptive L_1-norm penalty. The estimators have the oracle property Fan and Li (2001) when $p = o(n^{1/10})$. Hao et al. Hao et al. (2018) proposed a computationally efficient regularization algorithm under marginality principle (RAMP) that

simultaneously selects the main effects, interaction effects and quadratic effects for high-dimensional data $p \gg n$. They also verified the interaction selection consistency property of the two-stage LASSO under some sensible conditions.

The stage-wise selection procedure first performs a main effect selection (by excluding the interaction terms) to reduce the dimension of variables and then carries out a joint selection on the reduced list of variables, which is computationally feasible and effective. For example, viewing the sliced inverse regression Li (1991) from a likelihood perspective, Jiang and Liu Jiang and Liu (2014) suggested a stage-wise variable selection algorithm (SIRI) via inverse regression, which is able to detect higher order interactions without any specific hierarchical structure. Hao and Zhang Hao and Zhang (2014) proposed two stage-wise interaction selection procedures, IFORT and IFORM, both of which enjoy the sure screening property in the first stage. Fan et al. Fan et al. (2016) proposed a method, named the interaction pursuit, that incorporates both screening and variable selection in ultra-high dimensions. The method possesses both the sure screening property and the oracle property in the two stages respectively. For some other works on interaction selection, see Zhao et al. (2009); Li et al. (2012); Bien et al. (2013); Hall and Xue (2014). While having the aforementioned good properties, both types of interaction selection approaches have their own disadvantages as well. The joint selection is usually computational infeasible (insufficient storage) when p is large; the stage-wise selection, as pointed out in Hao and Zhang (2014), may be very difficult to be theoretically justified under general conditions.

Although there have been many novel developments on selection of interaction terms as described above, little work has been done on the estimation of the regression function when interactions exist. In this paper, we present some theoretical results on the minimax rate of convergence for estimating the high-dimensional regression function with interaction terms under three different hierarchical structures. Regardless of the heredity condition, our results show that the minimax rate is determined by the maximum of the total estimation price of the main effects and that of

the interaction effects. Heredity conditions enter the minimax rate of convergence in terms of the estimation price of the interaction effects, namely $r_2(1 + \log(K/r_2))/n$, where r_2 is the number of non-zero interaction effects and K is the number of eligible candidate interaction terms under each of the different heredity conditions. Consequently, a stronger heredity condition leads to possibly faster minimax rate of convergence. For example, when the underlying model has no more than r_1 non-zero main effects, at most $K = \binom{r_1}{2}$ interaction terms are allowed to enter the model under strong heredity, compared to $K = r_1(p_n - (r_1 + 1)/2)$ under weak heredity. As will be seen, only in certain situations is the minimax rate improved by imposing the strong heredity, although strong heredity allows fewer eligible interaction terms than the other two heredity conditions. Also, from the perspective of estimation, there may be no difference in rate of convergence between weak heredity and no heredity in many situations. An intuitive reason is that, when the number of interactions is small ($\log r_2$ is asymptotically away from $\log(r_1 p_n)$), the estimation price due to searching over the eligible interaction terms remains the same under the above two heredity conditions. Our results provide a complete characterization and comparison of the minimax rates of convergence under the three heredity conditions.

In real applications, since one does not know the true heredity condition behind the data (or practically the best heredity condition to describe the data at the given sample size), it is desirable to construct an estimator that performs optimally no matter which of the three heredity conditions holds. Such an estimator that adapts to the true heredity condition as well as the unknown number of main and interaction effects will be obtained in this paper.

The derivations of both the upper and lower bounds have close connections to the information theory. For the upper bound, the adoption of the model complexity in the model selection criterion used (ABC) is from the perspective of description length in information theory Rissanen (1983); Hansen and Yu (2001); Barron and Cover (1991); Wallace and Freeman (1987); Hall and Hannan (1988). The ABC criterion is inspired to handle the selection bias of AIC in high-dimensional case by adding an extra model

complexity term and it leads to desirable resolvability bounds. For the lower bound, Fano's inequality in information theory plays a key role.

The remainder of the paper is organized as follows. In Section 4.2, we introduce the model setup, the loss function and the heredity conditions for the problem. In Section 4.3, after stating the required assumption, we present our main results of the minimax rate of convergence under strong heredity. The theoretical results under weak heredity and no heredity are presented in Section 4.4. Section 4.5.1 provides detailed rates of convergence under different heredity conditions in relation to the sparsity indices, the ambient dimension and the sample size, followed by Section 4.5.2 where we present some interesting implications of the detailed results. In Section 4.6, we extend our results to quadratic models in which both quadratic and interaction effects are considered. In Section 4.7, we construct an adaptive estimator that achieves the minimax rate of convergence without knowledge of the type of the heredity condition or the sparsity indices (r_1 and r_2). The proofs of our results and some technical tools are presented in the Appendix.

4.2 Preliminaries

Model Setup Suppose the dataset is composed of $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is a $n \times p$ matrix with n observations on p covariates and $\mathbf{Y} = (y_1, ..., y_n)^T$ is the response vector. We start by considering a linear regression model with both main effects and two-way interaction effects:

$$\mathbf{Y} \quad = \quad \mathbf{Z}\beta + \epsilon, \tag{4.1}$$

where $\beta = ((\beta^{(1)})^T, (\beta^{(2)})^T)^T$ is the overall coefficient vector, $\mathbf{Z} = (\mathbf{X}, [\mathbf{XX}]) \in \mathbb{R}^{n \times (\frac{p^2+p}{2})}$ is the full design matrix, and the random noise vector $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ with known σ. More specifically, $\beta^{(1)} \in \mathbb{R}^p$ and $\beta^{(2)} \in \mathbb{R}^{\binom{p}{2}}$ are the coefficients of the main effects and the two-way interaction effects respectively. Here we define

$[\mathbf{XX}] = (\mathbf{x}_1 \circ \mathbf{x}_2, ..., \mathbf{x}_1 \circ \mathbf{x}_p, ..., \mathbf{x}_{p-1} \circ \mathbf{x}_p)^T$ as the $n \times \binom{p}{2}$ matrix that contains all the two-way interaction terms, where $\circ$ denotes the point-wise product of two vectors.

In this paper, our focus is on the fixed design, i.e., the covariates are considered given. Our goal is to estimate the mean regression function by a linear combination of the covariates and interaction terms.

Loss Function Denote $h(\cdot) : \mathbb{R}^{(p^2+p)/2} \to \mathbb{R}$ as the mean regression function, i.e., $h(\mathbf{z}) = \mathbf{z}^T\beta$ for $\mathbf{z} \in \mathbb{R}^{(p^2+p)/2}$. Denote $\hat{h}(\mathbf{z}) = \mathbf{z}^T\hat{\beta}$ as an estimated function of $h(\mathbf{z})$. In our fixed design setting, we focus on the prediction loss (or the Averaged Squared Error) $\mathsf{L}(h, \hat{h}) := \frac{1}{n}||\mathbf{Z}\beta - \mathbf{Z}\hat{\beta}||_2^2$, where $\|\cdot\|_2$ is the Euclidean norm. Set the index sets for the main effects and the interaction effects as $\mathtt{I}_{\mathtt{main}} = \{1, ..., p\}$ and $\mathtt{I}_{\mathtt{int}} = \{(i, j) : 1 \le i < j \le p\}$ respectively.

Let $\mathtt{I} = (\mathtt{I}_1, \mathtt{I}_2) \subset \mathtt{I}_{\mathtt{main}} \otimes \mathtt{I}_{\mathtt{int}}$ ($\otimes$ is the Cartesian product) be the index set of a model with $|\mathtt{I}_1|$ non-zero main effects and $|\mathtt{I}_2|$ non-zero interaction effects. In this paper, we consider the data generating model ((4.1)) with at least two main effects and one interaction effect purely for convenience, which does not affect the conclusions. Let $\mathbf{Z}_\mathtt{I}$ be the $n \times |\mathtt{I}|$ submatrix of $\mathbf{Z}$ that corresponds to the model index $\mathtt{I}$. Its corresponding least squares estimator $P_\mathtt{I}\mathbf{Y}$ is used to estimate $\mathbf{Z}\beta$, where $P_\mathtt{I}$ is the projection matrix onto the column space of $\mathbf{Z}_\mathtt{I}$. The loss function of using model $\mathtt{I}$ is denoted as $\mathcal{L}(\mathtt{I}) := \frac{1}{n}||P_\mathtt{I}\mathbf{Y} - \mathbf{Z}\beta||_2^2$.

Heredity Conditions Denote the space of all the $p + \binom{p}{2}$-dimensional vectors with a hierarchical notation of the subscripts as

$$\ddot{\mathbb{R}}^p = \{\beta \in \mathbb{R}^{p+\binom{p}{2}} | \beta = (\beta_1, ..., \beta_p, \beta_{1,2}, ..., \beta_{p-1,p})\}.$$

We refer to $\beta^{(1)} = (\beta_1, ..., \beta_p)$ as the subvector consisting of the first p elements in β, and $\beta^{(2)} = (\beta_{1,2}, ..., \beta_{p-1,p})$ as the subvector containing the rest of the elements. We

introduce the following two vector spaces:

$$\ddot{\mathbb{R}}^p_{weak} = \left\{\beta \in \ddot{\mathbb{R}}^p | \mathbb{1}_{\beta_{i,j}\neq 0} \leq \mathbb{1}_{\beta_i \neq 0} \vee \mathbb{1}_{\beta_j \neq 0}, 1 \leq i < j \leq p\right\}$$

and

$$\ddot{\mathbb{R}}^p_{strong} = \left\{\beta \in \ddot{\mathbb{R}}^p | \mathbb{1}_{\beta_{i,j}\neq 0} \leq \mathbb{1}_{\beta_i \neq 0} \cdot \mathbb{1}_{\beta_j \neq 0}, 1 \leq i < j \leq p\right\}.$$

The space $\ddot{\mathbb{R}}^p_{strong}$ captures the strong heredity condition that if the interaction term is in the model, then both of its corresponding main effects should also be included. The space $\ddot{\mathbb{R}}^p_{weak}$ characterizes the weak heredity condition that if the interaction is in the model, then at least one of its main effects should be included. As pointed out in Hao and Zhang (2017), the sign of the main effect coefficients are not invariant of linear transformation of the covariates individually due to the existence of the interaction terms. Heredity conditions are consequently meaningless without the specification of the model parametrization. In our paper, we stick to the parameterization $\mathbf{Z}$ and include the no heredity condition by considering the vector space $\ddot{\mathbb{R}}^p$. Define the l_0-norm of a vector $a = (a_1, ..., a_p)$ as the number of its non-zero elements, i.e., $\|a\|_0 = \sum_{i=1}^p \mathbb{1}_{a_i \neq 0}$. For a vector space $\mathcal{S} \in \left\{\ddot{\mathbb{R}}^p_{strong}, \ddot{\mathbb{R}}^p_{weak}, \ddot{\mathbb{R}}^p\right\}$, define the corresponding *l_0-ball* and *l_0-hull* of $\mathcal{S}$ as

$$B_0(r_1, r_2; \mathcal{S}) = \left\{\beta = (\beta^{(1)}, \beta^{(2)}) \in \mathcal{S}, \left\|\beta^{(1)}\right\|_0 \leq r_1, \left\|\beta^{(2)}\right\|_0 \leq r_2\right\} \tag{4.2}$$

and

$$\mathcal{F}_0(r_1, r_2; \mathcal{S}) = \left\{h : h(\mathbf{z}) = \mathbf{z}^T\beta, \beta \in B_0(r_1, r_2; \mathcal{S})\right\}$$

respectively. Note that $B_0(r_1, r_2; \mathcal{S})$ represents the collection of coefficients β with at most r_1 non-zero main effects and r_2 non-zero interaction effects under a certain hierarchical constraint $\mathcal{S}$. And $\mathcal{F}_0(r_1, r_2; \mathcal{S})$ denotes the collection of linear combinations of the covariates with coefficients $\beta \in B_0(r_1, r_2; \mathcal{S})$. Throughout this paper, we assume that $r_1 + r_2 \leq n$ (otherwise the minimax risk may not converge), $r_1 \geq 2$ and $r_2 \geq 1$.

Minimax Risk It is helpful to consider the uniform performance of a modeling procedure when we have plentiful choices of modeling procedures during the analysis of a statistical problem. The minimax framework seeks an estimator that minimizes the worst performance (in statistical risk) assuming that the truth belongs to a function class $\mathcal{W}$. The minimax risk we consider is

$$\min_{\hat{h}} \max_{h \in \mathcal{W}} E\mathsf{L}(\hat{h}, h),$$

where $\hat{h}$ is over all estimators, and min and max may refer to inf and sup, more formally speaking. In our work, we assume that the true mean regression function has a hierarchical structure by imposing $\mathcal{W} = \mathcal{F}_0(r_1, r_2; \mathcal{S})$, with $\mathcal{S} \in \left\{ \ddot{\mathbb{R}}^p_{strong}, \ddot{\mathbb{R}}^p_{weak}, \ddot{\mathbb{R}}^p \right\}$.

In this paper, we will use the notation $b_n \succeq a_n$ or $a_n \preceq b_n$ to represent $a_n = O(b_n)$. If both $b_n \succeq a_n$ and $a_n \succeq b_n$ hold, we denote $a_n \asymp b_n$ to indicate that a_n and b_n are of the same order. If $a_n \succeq b_n$ holds without $a_n \asymp b_n$, we use the notation $a_n \succ b_n$ or $b_n \prec a_n$.

4.3 Minimax Rate of Convergence under Strong Heredity

4.3.1 Assumption

We start by stating an assumption required for our result of the minimax rate of convergence under strong heredity. In this paper, we use p_n to indicate that the number of main effects p can go to infinity as n increases. We also allow r_1 and r_2 to increase with the sample size n as well.

Sparse Reisz Condition (SRC) For some $l_1, l_2 > 0$, there exist constants $b_1, b_2 > 0$ (not depending on n) such that for any $\beta = (\beta^{(1)}, \beta^{(2)})$ with $\left\|\beta^{(1)}\right\|_0 \leq \min(2l_1, \mathrm{p}_n)$

and $\left\|\beta^{(2)}\right\|_0 \leq \min\left(2l_2, \binom{p_n}{2}\right)$, we have

$$b_1 \|\beta\|_2 \leq \frac{1}{\sqrt{n}} \|\mathbf{Z}\beta\|_2 \leq b_2 \|\beta\|_2 . \tag{4.3}$$

The SRC assumption requires that the eigenvalues of $\frac{1}{n}\mathbf{Z}_{\mathtt{I}}^T\mathbf{Z}_{\mathtt{I}}$ for any relevant sparse submatrix $\mathbf{Z}_{\mathtt{I}}$ of $\mathbf{Z}$ are bounded above and away from 0. It was first proposed in Zhang and Huang (2008). It is similar to the sparse eigenvalue conditions in Zhang (2010c); Raskutti et al. (2011), quasi-isometry condition in Rigollet and Tsybakov (2011); it is also related to the more stringent restricted isometry property (which requires the constants b_1, b_2 are close to 1) in Candes and Tao (2007). Such assumptions are standard in the l_1-regularization analysis like LASSO and the Dantzig selector. See Bickel et al. (2009); Meinshausen and Yu (2009); Koltchinskii (2009) for more references.

One way to interpret the imposition of the SRC assumption is that $\|\theta - \beta\|_2^2$ characterizes, up to a constant, the Kullback-Leibler divergence between two joint densities (the joint distribution of the response vector y under fixed design) with parameters θ and β respectively, when θ and β are properly sparse. To see this, let $\mathbf{z}_i$ be the i-th row of $\mathbf{Z}$ and we have the joint density $P_\theta = (2\pi)^{-n/2}\sigma^{-n}\prod_{i=1}^n \exp(-\frac{1}{2}(y_i - \mathbf{z}_i\theta)^2/\sigma^2)$ with parameter θ. The K-L distance is then $D(P_\theta||P_\beta) = \frac{1}{2\sigma^2}\sum_{i=1}^n(\mathbf{z}_i\beta - \mathbf{z}_i\theta)^2 = \frac{1}{2\sigma^2}\|\mathbf{Z}(\theta-\beta)\|_2^2$, which behaves like $\|\theta-\beta\|_2^2$ under SRC .

Such a relationship between the regression function space and the coefficient space is needed in deriving the minimax lower bound. Without this assumption, the metric entropy of the regression function class may not be determined in terms of the numbers of the main and interaction terms, and the actual minimax risk can converge at different rates, depending on how $\|\mathbf{Z}(\theta-\beta)\|_2^2$ and $\|\theta-\beta\|_2^2$ are related. The SRC is a relatively mild condition that imposes constraints on the sub-matrices of $\mathbf{Z}$ with small sizes. It does not necessarily ensure that the design matrix has rank close to $\min(n, p_n)$. The SRC condition is expected to hold when the true regression function has a sparse representation and the covariates are not highly correlated.

4.3.2 Minimax rate

Now we present our main result of the minimax rate of convergence under strong heredity. A simple estimator is enough for an effective minimax upper bound. Let $\hat{\mathtt{I}} = \arg\min_{\mathtt{I}\in\mathtt{I}^{strong}_{r_1,r_2}} \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{\mathtt{I}})^2$ be the model that minimizes the residual sum of squares over all the models that have exactly r_1 non-zero main effects and r_2 non-zero interaction effects under strong heredity, denoted as $\mathtt{I}^{strong}_{r_1,r_2}$, where $\hat{\mathbf{Y}}^{\mathtt{I}} = P_{\mathtt{I}}\mathbf{Y}$ is the projection of $\mathbf{Y}$ onto the column space of the design matrix $\mathbf{Z}_{\mathtt{I}}$. For lower bounding the minimax risk, the information-theoretical tool of using Fano's inequality with metric entropy understanding Yang and Barron (1999) plays an important role in the proof.

Theorem 7 Under the Sparse Reisz Condition with $l_1 = r_1 \le p_n \wedge n$, $l_2 = r_2 \le \binom{r_1}{2} \wedge n$ and the strong heredity condition $\mathcal{W} = \mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})$, the minimax risk is upper bounded by

$$\min_{\hat{h}} \max_{h\in\mathcal{W}} E\mathsf{L}(\hat{h}, h) \le \sup_{h\in\mathcal{W}} E(\mathcal{L}(\hat{\mathtt{I}})) \le \frac{c\sigma^2}{n}\left(r_1\left(1 + \log\frac{p_n}{r_1}\right) + r_2\left(1 + \log\frac{\binom{r_1}{2}}{r_2}\right)\right), \tag{4.4}$$

where c is a pure constant; the minimax risk is lower bounded by

$$\min_{\hat{h}} \max_{h\in\mathcal{W}} E\mathsf{L}(\hat{h}, h) \ge c_1\frac{\sigma^2}{n}\left(r_1\left(1 + \log\frac{p_n}{r_1}\right) \vee r_2\left(1 + \log\frac{\binom{r_1}{2}}{r_2}\right)\right) \tag{4.5}$$

for some positive constant c_1 that only depends on the constants b_1 and b_2 in the SRC assumption. □

From the theorem, under the SRC and the strong heredity condition, the minimax rate of convergence scales as: $\min_{\hat{h}} \max_{h\in\mathcal{W}} E\mathsf{L}(\hat{h}, h) \asymp \frac{\sigma^2}{n}(r_1(1 + \log\frac{p_n}{r_1}) \vee r_2(1 + \log(\binom{r_1}{2}/r_2)))$.

Remark 14 The term $r_1(1+\log(p_n/r_1))/n = \frac{r_1}{n} + \frac{r_1}{n}\log(p_n/r_1)$ reflects two aspects in the estimation of the main effects: the price of searching among $\binom{p_n}{r_1}$ possible models, which is of order $r_1\log(p_n/r_1)/n$, and the price of estimating the r_1 main effect coefficients after the search. Thus $r_1(1+\log(p_n/r_1))/n$ is *the total price of estimating the main effects*. Similarly, $r_2\left(1+\log\left(\binom{r_1}{2}/r_2\right)\right)/n$ is *the total price of estimating the interaction effects*. □

Remark 15 Our result of the upper bound is general and holds regardless of the size of r_1. When r_1 is large (e.g., close to n), the upper bound converges slowly or even does not converge at all. □

4.4 Minimax Rate of Convergence under Weak Heredity and No Heredity

Similar results are obtained under weak heredity and no heredity. The minimax rate of convergence is still determined by the maximum of the total price of estimating the main effects and that of the interaction effects. When the heredity condition changes, the total price of estimating the interaction effects may differ, possibly substantially.

Theorem 8 Under the Sparse Reisz Condition with $l_1 = r_1 \le p_n \wedge n$, $l_2 = r_2 \le (r_1 p_n) \wedge n$ and the weak heredity condition $\mathcal{W} = \mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{weak})$, the minimax risk is of order

$$\min_{\hat{h}} \max_{h\in\mathcal{W}} E\mathsf{L}(\hat{h}, h) \asymp \frac{\sigma^2}{n}\left(r_1\left(1+\log\frac{p_n}{r_1}\right) \vee r_2\left(1+\log\frac{r_1\cdot p_n}{r_2}\right)\right). \tag{4.6}$$

□

Theorem 9 Under the Sparse Reisz Condition with $l_1 = r_1 \le p_n \wedge n$, $l_2 = r_2 \le$

$\binom{p_n}{2} \wedge n$ and the no heredity condition $\mathcal{W} = \mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n})$, the minimax risk is of order

$$\min_{\hat{h}} \max_{h \in \mathcal{W}} E\mathsf{L}(\hat{h}, h) \asymp \frac{\sigma^2}{n} \left(r_1 \left(1 + \log \frac{p_n}{r_1} \right) \vee r_2 \left(1 + \log \frac{\binom{p_n}{2}}{r_2} \right) \right). \tag{4.7}$$

□

Remark 16 We apply standard analytical tools in the derivations of minimax upper and lower bounds in the preceding theorems. For the *upper bound*, it is crucial to deal with the selection bias, which arises from the difficulty in identifying the set of nonzero coefficients among combinatorial many choices and thus can be very large since p_n is allowed to be arbitrarily large. Note that the familiar analyses and results for bias-correction type of criteria such as AIC are not applicable here. The oracle inequality for the ABC criterion turns out to work effectively with carefully designed model complexity terms for establishing the optimal-rate upper bounds. For the *lower bound*, Fano's inequity is expected to do the job, but there are significant details to work out to obtain matching upper and lower bounds in order. In particular, we need to sort out the metric entropy behaviors of the target function classes defined under the different heredity conditions, which involves the relationship between the parameter (coefficient) space and the regression function space. The risk bounds in the form of the maximum value of the precisely derived prices of estimating the main effects and the interactions respectively shed light on understanding how the number of the main effects and that of the interactions, together with the hierarchical structure, jointly determine the minimax rate of convergence. □

4.5 Comparisons and Insights

In this section, we summarize the consequences of our main results in three scenarios for an integrated understanding. For brevity, we introduce the following notation. For $a, b \in \mathbb{N}^+$ and $a \geq b$, define the quantity $\xi_b^a := b(1 + \log(a/b))$. The total price of

estimating the main effects and the interaction effects are then denoted as $\sigma^2\xi_{r_1}^{p_n}/n$ and $\sigma^2\xi_{r_2}^{K}/n$ respectively, where K depends on p_n, r_1 and the heredity condition. We also use the notation $K_{\mathcal{S}}$ ((4.14)) to indicate that K depends on the heredity condition $\mathcal{S}$. Let

$$\mathcal{M}(\mathcal{S}) := \min_{\hat{h}} \max_{h\in\mathcal{F}_0(r_1,r_2;\mathcal{S})} E\mathsf{L}(\hat{h}, h)$$

denote the minimax risk under the heredity condition $\mathcal{S}$.

4.5.1 Detailed rates of convergence

Since the minimax rate of convergence depends on the maximum of $\xi_{r_1}^{p_n}$ and $\xi_{r_2}^{K}$, we discuss the cases where one of the two quantities is greater than the other.

Scenario 1: $r_2 \preceq r_1$ When there are more main effects than interaction effects in the sense that $r_2 \preceq r_1$, the minimax rate of convergence is not affected by the heredity conditions. When $\log(p_n/r_1) \succeq \log r_1$, i.e., $\log(p_n/r_1) \asymp \log p_n$, we always have $\xi_{r_1}^{p_n} \succeq \xi_{r_2}^{p_n^2} = \max\{\xi_{r_2}^{r_1^2}, \xi_{r_2}^{r_1 p_n}, \xi_{r_2}^{p_n^2}\}$, i.e., $\xi_{r_1}^{p_n} \succeq \xi_{r_2}^{K}$ regardless of the heredity conditions. When $\log(p_n/r_1) \prec \log r_1$, it depends on the order of r_2 to further decide which estimation price is larger. When $\log(p_n/r_1) \prec \log r_1$, let r_* be such that $\xi_{r_1}^{p_n} \asymp \xi_{r_*}^{r_1^2}$. If $r_* \succeq r_2$, we have $\xi_{r_1}^{p_n} \succeq \xi_{r_2}^{K}$; otherwise $\xi_{r_1}^{p_n} \prec \xi_{r_2}^{K}$.

In summary, given that $r_2 \preceq r_1$, the minimax risk is of order

$$\mathcal{M}(\mathcal{S}) \asymp \begin{cases} \frac{\sigma^2}{n}\xi_{r_2}^{r_1^2}, & \text{if } r_* \preceq r_2 \preceq r_1 \text{ and } \log\frac{p_n}{r_1} \prec \log r_1, \\ \frac{\sigma^2}{n}\xi_{r_1}^{p_n}, & \text{otherwise,} \end{cases}$$

for $\mathcal{S} \in \left\{\ddot{\mathbb{R}}^p_{strong}, \ddot{\mathbb{R}}^p_{weak}, \ddot{\mathbb{R}}^p\right\}$.

Remark 17 The cutoff relationship $\log(p_n/r_1) \succeq \log r_1$, or equivalently $\log(p_n/r_1) \asymp \log p_n$, actually characterizes the sparseness of the main effects. It requires sparseness in log order that $\log r_1$ is not too close to $\log p_n$. For example, $\log(p_n/r_1) \succeq \log r_1$ holds

when $p_n \asymp \exp(r_1)$ or $p_n \asymp r_1^{1+\alpha}$ with a constant $\alpha > 0$, but not when $p_n \asymp r_1 \log(r_1)$, although these cases all satisfy that $r_1 \ll p_n$. More insights of Scenario 1 are discussed in 2) of subsection 4.5.2. □

Remark 18 This scenario also includes the special case when $p_n = O(1)$, where we must have $r_1 = O(1)$ and $r_2 = O(1)$. The minimax rate of convergence is of the standard parametric order $1/n$ regardless of the heredity conditions. □

Scenario 2: $r_1 \preceq r_2$ and $\log p_n \preceq r_1$ When there exist more interaction terms, i.e., $r_1 \preceq r_2$, under weak or no heredity, the quantity $\xi_{r_2}^K$ is always no less than (in order) $\xi_{r_1}^{p_n}$.

For strong heredity, we discuss case by case. When $\log(p_n/r_1) \prec \log r_1$, we always have $\xi_{r_1}^{p_n} \preceq \xi_{r_2}^{r_1^2}$. When $\log(p_n/r_1) \succeq \log r_1$, it depends on the order of r_2 to decide which estimation price is larger in terms of order. When $\log(p_n/r_1) \succeq \log r_1$, let r'_* be such that $\xi_{r_1}^{p_n} \asymp \xi_{r'_*}^{r_1^2}$. If $r_2 \succeq r'_*$, we have $\xi_{r_1}^{p_n} \preceq \xi_{r_2}^{r_1^2}$; otherwise $\xi_{r_1}^{p_n} \succ \xi_{r_2}^{r_1^2}$. In summary, given that $r_1 \preceq r_2$ and $\log p_n \preceq r_1$, the minimax risk is of order

$$\mathcal{M}(\ddot{\mathbb{R}}_{strong}^{p_n}) \asymp \begin{cases} \frac{\sigma^2}{n}\xi_{r_1}^{p_n}, & \text{if } r_1 \preceq r_2 \preceq r'_* \text{ and } \log\frac{p_n}{r_1} \succeq \log r_1, \\ \frac{\sigma^2}{n}\xi_{(r_2 \wedge r_1^2)}^{r_1^2}, & \text{otherwise,} \end{cases}$$

$$\mathcal{M}(\ddot{\mathbb{R}}_{weak}^{p_n}) \asymp \frac{\sigma^2}{n}\xi_{(r_2 \wedge r_1 p_n)}^{r_1 p_n},$$

$$\mathcal{M}(\ddot{\mathbb{R}}^{p_n}) \asymp \frac{\sigma^2}{n}\xi_{(r_2)}^{p_n^2}.$$

Remark 19 The term $\xi_{(r_2 \wedge K)}^K$ deals with the case where r_2 is inactive in the sense that r_2 exceeds K under the specific heredity condition. For example, with $r_2 \geq \binom{r_1}{2}$, the upper bound r_2 in ((4.2)) does not provide any new information of the number of non-zero interaction effects for strong heredity. Thus the l_0-ball $B_0(r_1, r_2; \ddot{\mathbb{R}}_{strong}^p)$ is automatically reduced to a subset $B_0(r_1, \binom{r_1}{2}; \ddot{\mathbb{R}}_{strong}^p)$. □

Scenario 3: $r_1 \preceq r_2$ and $\log p_n \succeq r_1$ When the number of the main effects p_n is at least exponentially as many as the non-zero main effects in the sense that $\log p_n \succeq r_1$, $\xi_{r_1}^{p_n}$ is always no less than $\xi_{r_2}^{K}$ in terms of order. In fact, in this scenario, the results of the minimax rates under weak or no heredity are exactly the same as those in Scenario 2. For completeness, we still present the results. Specifically, the minimax risk is of order

$$\mathcal{M}(\ddot{\mathbb{R}}_{strong}^{p_n}) \asymp \frac{\sigma^2}{n}\xi_{r_1}^{p_n},$$

$$\mathcal{M}(\ddot{\mathbb{R}}_{weak}^{p_n}) \asymp \frac{\sigma^2}{n}\xi_{(r_2 \wedge r_1 p_n)}^{r_1 p_n},$$

$$\mathcal{M}(\ddot{\mathbb{R}}^{p_n}) \asymp \frac{\sigma^2}{n}\xi_{r_2}^{p_n^2}.$$

4.5.2 Interesting implications

1. Comparing the results for weak heredity and no heredity, we may or may not have distinct rates of convergence. When there exists a small constant $c > 0$ such that $\log r_2 \leq (1-c) \cdot \log(r_1 p_n)$ for large enough n, there is no difference between weak heredity and no heredity from the perspective of rate of convergence in estimation. It still remains an open question how they are different for the problem of model identification. Without the above relationship between r_1 and r_2, there is no guarantee that the rates of convergence are the same under weak heredity and no heredity. For example, when $r_2 = r_1 p_n / \log r_1$, if in addition we have $r_1 = p_n \leq n^{1/2}$, the minimax rates are the same under weak and no heredity, at $\mathcal{M}(\ddot{\mathbb{R}}_{weak}^{p_n}) \asymp \mathcal{M}(\ddot{\mathbb{R}}^{p_n}) \asymp r_1 p_n \log\log r_1/(n \log r_1)$. In contrast, if instead we have $r_1 = \sqrt{p_n}$, then the minimax rates are different, with $\mathcal{M}(\ddot{\mathbb{R}}_{weak}^{p_n}) \asymp r_1 p_n \log\log r_1/(n \log r_1)$ and $\mathcal{M}(\ddot{\mathbb{R}}^{p_n}) \asymp r_1 p_n/n$.

2. Heredity conditions do not affect the rates of convergence in some situations. For example, when there exist more main effects than interaction effects (Scenario 1), the minimax rates of convergence are the same under all three heredity

conditions. To understand why the heredity condition is blurred when the number of main effects dominates, we first observe the risk increment from strong heredity ($(\frac{\sigma^2}{n}(\xi_{r_1}^{p_n} + \xi_{r_2}^{r_1^2}))$ to no heredity ($(\frac{\sigma^2}{n}(\xi_{r_1}^{p_n} + \xi_{r_2}^{p_n^2}))$. Note the risk bound increment is of order $2\frac{r_2}{n}\log\frac{p_n}{r_1}$, which is smaller than $2\frac{r_1}{n}(1+\log\frac{p_n}{r_1}) \asymp \frac{1}{n}\xi_{r_1}^{p_n}$ when $r_2 \preceq r_1$. Thus, the risk increment does not affect the rate of the convergence. The estimation price of the interaction terms may be of a higher order than that of the main terms, but interestingly in this case ($r_2 \preceq r_1$), the differences among the prices in learning the interaction terms under different heredity conditions are always not larger (in order) than the price of learning the main effects.

3. From the detailed rates of convergence, under any of the three heredity conditions, the estimation of the interaction terms $\xi_{r_2}^K/n$ may become the dominating part. There are two different reasons why the price of estimating the interaction terms becomes higher than that for the main effect terms. One is that the number of interaction terms is more than that of the main effect terms. The other reason is that although the main effect terms outnumber the interaction terms, the ambient dimension is so large that even estimating a small number of the interaction terms is more challenging than estimating the main effects.

4. How much can the rate of convergence be improved by imposing strong heredity? We quantify this improvement by taking the ratio of two minimax rates of convergence given the ambient dimension p_n, i.e., $\mathcal{M}(\ddot{\mathbb{R}}_{strong}^{p_n})/\mathcal{M}(\ddot{\mathbb{R}}_{weak}^{p_n})$ and $\mathcal{M}(\ddot{\mathbb{R}}_{strong}^{p_n})/\mathcal{M}(\ddot{\mathbb{R}}^{p_n})$. In Scenario 2 ($r_1 \preceq r_2$ and $\log p_n \preceq r_1$), we have $\mathcal{M}(\ddot{\mathbb{R}}_{strong}^{p_n})/\mathcal{M}(\ddot{\mathbb{R}}_{weak}^{p_n}) \succeq \log p_n/p_n$, where the maximal improvement happens when $r_1 \asymp \log p_n$ and $r_2 \asymp r_1 p_n$. That is, the minimax rate of convergence under strong heredity is up to $\log p_n/p_n$ times faster than that under weak heredity. Similarly we have $\mathcal{M}(\ddot{\mathbb{R}}_{strong}^{p_n})/\mathcal{M}(\ddot{\mathbb{R}}^{p_n}) \succeq \log^2 p_n/p_n^2$, where the maximal improvement $\log^2 p_n/p_n^2$ happens at $r_1 \asymp \log p_n$ and $r_2 \asymp p_n^2$.

5. In Scenario 3 ($r_1 \preceq r_2$ and $\log p_n \succeq r_1$), the improvement

$\mathcal{M}(\ddot{\mathbb{R}}^{p_n}_{strong})/\mathcal{M}(\ddot{\mathbb{R}}^{p_n}_{weak}) \succeq \log p_n/p_n$, where the maximal improvement happens when $r_2 \succeq r_1 p_n$. In this scenario, the maximal improvement of the minimax rate from weak heredity to strong heredity depends on the ambient dimension p_n. In other words, the larger the ambient dimension is, the more improvement of minimax rate of convergence we have from weak heredity to strong heredity. Similarly we have $\mathcal{M}(\ddot{\mathbb{R}}^{p_n}_{strong})/\mathcal{M}(\ddot{\mathbb{R}}^{p_n}) \succeq \log p_n/p_n^2$, where the equality holds if $r_1 = O(1)$ and $r_2 \asymp p_n^2$.

6. If r_2 is active for all three heredity conditions, i.e., $r_2 \leq \binom{r_1}{2}$, the maximal improvement of minimax rate from weak/no heredity to strong heredity turns out to be consistent. That is, $\mathcal{M}(\ddot{\mathbb{R}}^{p_n}_{strong})/\mathcal{M}(\ddot{\mathbb{R}}^{p_n}_{weak}) \asymp \mathcal{M}(\ddot{\mathbb{R}}^{p_n}_{strong})/\mathcal{M}(\ddot{\mathbb{R}}^{p_n}) \succeq 1/\log p_n$, where the maximal improvement happens at $r_1 \asymp \log p_n$ and $r_2 \asymp r_1^2$.

4.6 Extension to Quadratic Models

Our aforementioned results do not consider quadratic effects. When both quadratic and two-way interaction effects are included in a model (called a quadratic model), it is easy to see the rates of convergence in the theorems still apply under both strong heredity and weak heredity. However, in the case of no heredity, the number of quadratic terms enters into the minimax rate. Assume one model has at most r_3 extra non-zero quadratic terms. We need the following assumption.

Sparse Reisz Condition 2 (SRC2) For some $l_1, l_2, l_3 > 0$, there exist constants $b_1, b_2 > 0$ (not depending on n) such that for any $\beta = (\beta^{(1)}, \beta^{(2)}, \beta^{(3)})$ with $\left\|\beta^{(1)}\right\|_0 \leq \min(2l_1, \mathrm{p}_n)$, $\left\|\beta^{(2)}\right\|_0 \leq \min(2l_2, \binom{p_n}{2})$ and $\left\|\beta^{(3)}\right\|_0 \leq \min(2l_3, p_n)$, we have

$$b_1 \|\beta\|_2 \leq \frac{1}{\sqrt{n}} \|\mathbf{Z}^*\beta\|_2 \leq b_2 \|\beta\|_2 ,$$

where $\mathbf{Z}^* = (\mathbf{X}, [\mathbf{XX}], \mathbf{X^2})$ is the new design matrix, with $\mathbf{X^2}$ representing the $n \times p$ matrix that contains all the quadratic terms.

Next we state the minimax results for quadratic models. Strong heredity and weak heredity are exactly the same condition since a quadratic term has only one corresponding main effect term. That is, both strong and weak heredity require that if a quadratic term X_1^2 has a non-zero coefficient, then X_1 must also have a non-zero coefficient. Similarly, under SRC2 with $l_1 = r_1, l_2 = r_2, l_3 = r_3$, the minimax rate of convergence under strong/weak heredity for the quadratic model stays the order

$$\frac{\sigma^2}{n}\left(r_1(1+\log\frac{p_n}{r_1}) \vee r_2(1+\log\frac{\binom{r_1}{2}}{r_2})\right); \tag{4.8}$$

under no heredity, its order becomes

$$\frac{\sigma^2}{n}\left(\bar{r}(1+\log\frac{p_n}{\bar{r}}) \vee r_2(1+\log\frac{\binom{p_n}{2}}{r_2})\right), \tag{4.9}$$

where $\bar{r} = r_1 \vee r_3$.

Remark 20 The proofs of the rates are similar with the proofs in the two-way interaction case. So we do not include them in the paper. □

4.7 Adaptation to Heredity Conditions and Sparsity Indices

In the previous sections, we have determined the minimax rates of convergence for estimating the linear regression function with interactions under different sizes of sparsity indices r_1, r_2 and heredity conditions $\mathcal{S}$. These results assume that r_1, r_2 and $\mathcal{S}$ are known. However, in practice, we usually have no prior information about the underlying heredity condition nor the sparsity constraints. Thus it is necessary and appealing to build an estimator that adaptively achieves the minimax rate of

convergence without the knowledge of $\mathcal{S}$, r_1 and r_2. We construct such an adaptive estimator as below.

To achieve our goal, we consider one specific model and three types of models together as the candidate models:

$$\bar{\mathcal{F}} = \{\mathtt{I}_{p_n,(p_n^2-p_n)/2}\} \cup \{\mathtt{I}_{k_1,k_2}^{strong}\} \cup \{\mathtt{I}_{k_1,k_2}^{weak}\} \cup \{\mathtt{I}_{k_1,k_2}^{no}\},$$

where $\mathtt{I}_{p_n,(p_n^2-p_n)/2}$ denotes the full model with p_n main effects and all the $\binom{p_n}{2}$ interaction effects. It is included so that the risk of our estimator will not be worse than order $R_{\mathbf{Z}}/n$, in which $R_{\mathbf{Z}}$ is the rank of the full design matrix. With a slight abuse of the notation, we use $\mathtt{I}_{k_1,k_2}^{strong}$, $\mathtt{I}_{k_1,k_2}^{weak}$ and $\mathtt{I}_{k_1,k_2}^{no}$ to represent a model with k_1 main effects and k_2 interaction effects under strong heredity, weak heredity and no heredity respectively. Note that some models appear more than once in $\bar{\mathcal{F}}$, which does not cause any problem for the goal of estimating the regression function. The details of the range of k_1 and k_2 for each model class are shown in ((4.11)), ((4.12)) and ((4.13)).

Model selection criteria with a bias-correction term (e.g., AIC Akaike (1974), FPE Akaike (1969), C_p Mallows (1973)) have been studied and shown to have asymptotic optimal properties (e.g., Shibata (1983); Li (1987); Polyak and B. Tsybakov (1990)) under the constraint that there are only polynomially many models per size in the candidate set. However, when an exponential number of models or more are considered in high-dimensional cases, these selection criteria may fail due to severe selection bias (see Yang and Barron (1998)). The ABC criterion in Yang (1999) was proposed to overcome this limitation by adding an extra model complexity term. The selected model by ABC was proved to have desirable resolvability bounds. So we apply the ABC criterion to select the best model from the candidate set. Note that ABC was derived under information-theoretic considerations, heavily influenced by works in the intersection of information theory and statistics Barron and Cover (1991); Barron et al. (1994); Yang and Barron (1998). The specific application of the general ABC

criterion for the present problem also naturally follows from a coding perspective.

For a model $\mathtt{I}$ in $\bar{\mathcal{F}}$, the criterion value is

$$ABC(\mathtt{I}) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{\mathtt{I}})^2 + 2r_{\mathtt{I}}\sigma^2 + \lambda\sigma^2 C_{\mathtt{I}}, \tag{4.10}$$

where $\hat{\mathbf{Y}}^{\mathtt{I}} = P_{\mathtt{I}}\mathbf{Y}$ is the projection of $\mathbf{Y}$ onto the column space of the design matrix $\mathbf{Z}_{\mathtt{I}}$ with rank $r_{\mathtt{I}}$, $C_{\mathtt{I}}$ is the descriptive complexity of model $\mathtt{I}$ and $\lambda > 0$ is a constant. The model descriptive complexity satisfies $C_{\mathtt{I}} > 0$ and $\sum_{\mathtt{I}\in\bar{\mathcal{F}}}\exp(-C_{\mathtt{I}}) \leq 1$. From an information-theoretic perspective, the model complexity term can be considered as the code-length of a prefix-code that describes the model.

The model descriptive complexity is crucial in building the adaptive model. Let $\pi_0, \pi_1, \pi_2, \pi_3 \in (0,1)$ be four constants such that $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$. Set $C_{\mathtt{I}_{p_n,(p_n^2-p_n)/2}} = -\log\pi_0$ for the full model,

$$C_{\mathtt{I}^{strong}_{k_1,k_2}} = -\log\pi_1 + \log(p_n \wedge n) + \log\left(\binom{k_1}{2} \wedge n\right) + \log\binom{p_n}{k_1} + \log\binom{\binom{k_1}{2}}{k_2} \tag{4.11}$$

for $1 \leq k_1 \leq p_n \wedge n$ and $0 \leq k_2 \leq \binom{k_1}{2} \wedge n$,

$$C_{\mathtt{I}^{weak}_{k_1,k_2}} = -\log\pi_2 + \log(p_n \wedge n) + \log(K \wedge n) + \log\binom{p_n}{k_1} + \log\binom{K}{k_2} \tag{4.12}$$

with $K = k_1 p_n - \binom{k_1}{2} - k_1$ for $1 \leq k_1 \leq p_n \wedge n$ and $0 \leq k_2 \leq K \wedge n$, and

$$C_{\mathtt{I}^{no}_{k_1,k_2}} = -\log\pi_3 + \log(p_n \wedge n) + \log\left(\binom{p_n}{2} \wedge n\right) + \log\binom{p_n}{k_1} + \log\binom{\binom{p_n}{2}}{k_2}, \tag{4.13}$$

for $1 \leq k_1 \leq p_n \wedge n$ and $0 \leq k_2 \leq \binom{p_n}{2} \wedge n$. This complexity assignment recognizes that there are three types of models under the different heredity conditions.

Let $\hat{\mathtt{I}} = \arg\min_{\mathtt{I}\in\bar{\mathcal{F}}} ABC(\mathtt{I})$ denote the model that minimizes the ABC criterion over the candidate model set $\bar{\mathcal{F}}$ and $\hat{\mathbf{Y}}^{\hat{\mathtt{I}}} := P_{\hat{\mathtt{I}}}\mathbf{Y}$ denote the least squares estimate of

$\mathbf{Y}$ using the model $\hat{\mathtt{I}}$. Then we have the following oracle inequality.

Theorem 10 When $\lambda \geq 5.1/\log 2$, the worst risk of the ABC estimator $\hat{\mathbf{Y}}^{\hat{\mathtt{I}}}$ is upper bounded by

$$\sup_{h \in \mathcal{F}_0(r_1, r_2; \mathcal{S})} E(\mathcal{L}(\hat{\mathtt{I}})) \leq \frac{c\sigma^2}{n} \left[R_{\mathbf{Z}} \wedge \left(r_1 \left(1 + \log \frac{p_n}{r_1} \right) + r_2 \left(1 + \log \frac{K_{\mathcal{S}}}{r_2} \right) \right) \right],$$

with

$$K_{\mathcal{S}} = \begin{cases} \binom{r_1}{2}, & \text{if } \mathcal{S} = \ddot{\mathbb{R}}^p_{strong}, \\ r_1 p_n, & \text{if } \mathcal{S} = \ddot{\mathbb{R}}^p_{weak}, \\ \binom{p_n}{2}, & \text{if } \mathcal{S} = \ddot{\mathbb{R}}^p, \end{cases} \tag{4.14}$$

where $R_{\mathbf{Z}}$ is the rank of the full design matrix $\mathbf{Z}$ and the constant c only depends on the constant λ. □

From the theorem, without any prior knowledge of the sparsity indices, the constructed ABC estimator adaptively achieves the minimax upper bound regardless of the heredity conditions. The result also indicates a major difference between estimation and model identification. For estimation, from the result, we are able to achieve adaptation with respect to the heredity condition without any additional assumption. For model identification, although we are not aware of any work that addresses the task of adaptation over the unknown heredity nature, it seems certain that much stronger assumptions than those for consistency under an individual heredity condition will be necessary to achieve adaptive selection consistency. Achieving adaptive model selection consistency under different types of conditions remains an important open problem on model selection theory and methodology.

Remark 21 We do not require any assumptions on the relationship among the variables for the upper bound in the theorem. In particular, the variables may be arbitrary correlated. □

Remark 22 The order $R_{\mathbf{Z}}/n$ is achievable when we use the projection estimator from the full model. Thus the minimax rate of convergence is no slower than the order $R_{\mathbf{Z}}/n$. As is known, the rank of the design matrix plays an important role in determining the minimax rate of convergence under fixed design Yang (1999); Rigollet and Tsybakov (2011); Wang et al. (2014). For our result, when p_n, r_1 and r_2 together make the total estimation price of the true model small enough, the upper bound will be improved from $R_{\mathbf{Z}}/n$ to $(r_1(1+\log(p_n/r_1)) \vee r_2(1+\log(\binom{r_1}{2}/r_2)))/n$. □

Remark 23 The ABC estimator may not be practical when p_n is large. In such a case, stochastic search instead of all subset selection may be used for implementation, although the associated theoretical understanding is yet to be established. □

Remark 24 The term "$R_{\mathbf{Z}} \wedge$ " automatically applies to the lower bound under whichever heredity condition, since under the SRC assumption, it intrinsically requires that $r_1(1+\log(p_n/r_1)) \vee r_2(1+\log(\binom{r_1}{2}/r_2))$ is no larger than $R_{\mathbf{Z}}$ in terms of order. Otherwise, the lower bound $(r_1(1+\log(p_n/r_1)) \vee r_2(1+\log(\binom{r_1}{2}/r_2)))/n$ by our proof will exceed the upper bound $R_{\mathbf{Z}}/n$, which leads to a contradiction. We give a specific example in Appendix C.5 to illustrate this requirement. □

Chapter 5

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 paper, 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.

The proposed CVIL takes the "prediction" importance of a variable into consideration and incorporates two ways of examining variable importance: position and replaceability. CVIL_p concerns that the change of the prediction performance if we only have a limited information of a variable (the mean), and CVIL_r focuses on if a variable could be replaced in the sense that the removal of a replaceable variable will not substantially affect the prediction performance. The combination of the two variable importances can also be a tool for the purpose of model identification. Further investigation are then desired for those replaceable variables rather simply looking into highly correlated variables, which saves more time and money. The definition of CVIL provides a new way to understand how a variable is important to the modeling procedure, especially when a model with good performance is hard to explain and interpret, which is common in practice. It is also worth pointing out that CVIL-δ improves the stability of the modeling procedure when δ is unstable. The unstableness of a modeling procedure sometimes results from the high correlation of the variables in the dataset. One limitation of our work is that the replaceability impor-

tance (CVIL_r) cannot determine which one of the two highly correlated variables is in the true model (assuming there is an underlying model). Another shortcoming is that CVIL_r is computationally demanding since it runs the procedure every time it deletes a variable. Future work of the above two aspects is of great interest.

For estimation problems with existence of interaction terms, it is of potential interest that which hierarchical structure should be used in practice. An potential topic is to develop some tests for heredity condition since we may not have enough data to utilize no heredity condition. Then whether using strong heredity has significant difference/improvement from using weak heredity condition is crucial for real problems in many applications. The computational cost in interaction selection and estimation problems plays a key role. Another potential problem is to design a genetic algorithm to realize the ABC model selection criterion which achieves the minimax rate of convergence.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413.

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054.

Barron, A. R., Yang, Y., and Yu, B. (1994). Asymptotically optimal function estimation by minimum complexity criteria. *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 38.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bleich, J., Kapelner, A., George, E. I., Jensen, S. T., et al. (2014). Variable selection for bart: an application to gene regulation. *The Annals of Applied Statistics*, 8(3):1750–1781.

Braun, M. T. and Oswald, F. L. (2011). Exploratory regression analysis: A tool for selecting models and determining predictor importance. *Behavior Research Methods*, 43(2):331–339.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, 53(2):603–618.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114(3):542.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Chambaz, A., Neuvial, P., and van der Laan, M. J. (2012). Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic journal of statistics*, 6:1059.

Chen, L., Giannakouros, P., and Yang, Y. (2007). Model combining in factorial data analysis. *Journal of Statistical Planning and Inference*, 137(9):2920–2934.

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(2):321–334.

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280–293.

Chevan, A. and Sutherland, M. (1991). Hierarchical partitioning. *The American Statistician*, 45(2):90–96.

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(16):6287–6292.

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 24(1):17–36.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364.

Cook, R. D. and Weisberg, S. (2009). *Applied regression including computing and graphics*, volume 488. John Wiley & Sons.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Ehrenberg, A. S. C. (1990). The unimportance of relative importance. *American Statistician*, 44(3):260–260.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, Y., Kong, Y., Li, D., and Lv, J. (2016). Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*.

Feldman, B. et al. (1999). The proportional value of a cooperative game. *Manuscript. Chicago: Scudder Kemper Investments*.

Feldman, B. E. (2005). Relative importance and value. *Available at SSRN 2255827*.

Ferrari, D. and Yang, Y. (2015). Confidence sets for model selection by F-testing. *Statistica Sinica*, 25:1637–1658.

Fisher, A., Rudin, C., and Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the" rashomon" perspective. *arXiv preprint arXiv:1801.01489*.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gromping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):137–152.

Grömping, U. et al. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, 17(1):1–27.

Hall, P. and Hannan, E. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika*, 75(4):705–714.

Hall, P. and Xue, J.-H. (2014). On selecting interacting features from high-dimensional data. *Computational Statistics & Data Analysis*, 71:694–708.

Hamada, M. and Wu, C. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24(3):130–137.

Han, L., Zhang, Y., Wan, X.-F., and Zhang, T. (2016). Generalized hierarchical sparse model for arbitrary-order interactive antigenic sites identification in flu virus data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 865–874. ACM.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Hansen, M. H. and Yu, B. (2001). Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association*, 96(454):746–774.

Hao, N., Feng, Y., and Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

Hao, N. and Zhang, H. H. (2017). A note on high-dimensional linear regression with interactions. *The American Statistician*, 71(4):291–297.

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(1):21–34.

Hejazi, N. S., Kherad-Pajouh, S., van der Laan, M. J., and Hubbard, A. E. (2017). Variance stabilization of targeted estimators of causal parameters in high-dimensional settings. *arXiv preprint arXiv:1710.05451*.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, 57(2):116.

Huang, J., Ma, S., and Zhang, C.-H. (2008a). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618.

Huang, L.-S., Chen, J., et al. (2008b). Analysis of variance, coefficient of determination and f-test for local polynomial regression. *The Annals of Statistics*, 36(5):2085–2109.

Ioannidis, J. P. and Khoury, M. J. (2011). Improving validation practices in "omics" research. *Science*, 334(6060):1230–1232.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537.

Jiang, B. and Liu, J. S. (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, 42(5):1751–1786.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828.

Kruskal, W. and Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician*, 43(1):2–6.

Lai, R. C., Hannig, J., and Lee, T. C. (2015). Generalized fiducial inference for ultrahigh-dimensional regression. *Journal of the American Statistical Association*, 110(510):760–772.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

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(450):547–560.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Liang, H., Zou, G., Wan, A. T., and Zhang, X. (2012). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Number 519.535 L743. Scott, Foresman.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, pages 431–439.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A):3498–3528.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4):661–675.

McNutt, M. (2014). Raising the bar. *Science*, 345(6192):9–9.

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

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270.

Nan, Y. and Yang, Y. (2014). Variable selection diagnostics measures for high-dimensional regression. *Journal of Computational and Graphical Statistics*, 23(3):636–656.

Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, 140(1):48–77.

Olden, J. D., Joy, M. K., and Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3):389 – 397.

Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41(4):311–313.

Polyak, B. and B. Tsybakov, A. (1990). Asymptotic optimality of the C_p test for the orthogonal series estimation of regression. *Theory of Probability and Its Applications*, 35:293–306.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771.

Rinaldo, A., Wasserman, L., G'Sell, M., and Lei, J. (2016). Bootstrapping and Sample Splitting For High-Dimensional, Assumption-Free Inference.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431.

Rolling, C. A. and Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):749–769.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Sandri, M. and Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17(3):611–628.

Sapp, S., van der Laan, M. J., and Page, K. (2014). Targeted estimation of binary variable importance measures with interval-censored outcomes. *The international journal of biostatistics*, 10(1):77–97.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, 35(3):415–423.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076–1083.

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2:1–19.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008a). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008b). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307.

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(1):25.

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(43):15545–15550.

Theil, H. and Chung, C. (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions. *The American Statistician*, 42(4):249–252.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

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(2):183.

Van De Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Van der Laan, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1).

Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(3):240–252.

Wang, M., Chen, X., and Zhang, H. (2010). Maximal conditional chi-square importance in random forests. *Bioinformatics*, 26(6):831–837.

Wang, Z., Paterlini, S., Gao, F., and Yang, Y. (2014). Adaptive minimax regression estimation over sparse ℓ_q-hulls. *Journal of Machine Learning Research*, 15(1):1675–1711.

Wei, P., Lu, Z., and Song, J. (2015). Variable importance analysis: a comprehensive review. *Reliability Engineering & System Safety*, 142:399–432.

Williamson, B. D., Gilbert, P. B., Simon, N., and Carone, M. (2017). Nonparametric variable importance assessment using machine learning techniques.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102(477):235–243.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499.

Yang, Y. (2000). Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica*, 10(4):1069–1090.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588.

Yang, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, 13:783–809.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599.

Yang, Y. and Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116.

Ye, C. and Yang, Y. (2019). High-dimensional adaptive minimax sparse estimation with interactions. *IEEE Transactions on Information Theory*.

Ye, C., Yang, Y., and Yang, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association*, 113(524):1797–1812.

Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738–1757.

Yuan, Z. and Ghosh, D. (2008). Combining multiple biomarker models in logistic regression. *Biometrics*, 64(2):431–439.

Zhang, C. (2010a). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. (2010b). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhang, T. (2010c). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107.

Zhang, X., Lu, Z., and Zou, G. (2013). Adaptively combined forecasting for discrete response time series. *Journal of Econometrics*, 176(1):80–91.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Appendix A

Proofs and Supplemental Materials of Chapter 2

A.1 Proof of Theorem 1

Proof A.1

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,

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

Hence,

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

On the other hand,

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

A.2 Proof of Theorem 2

Proof A.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 $P\left(\frac{|\overline{\mathcal{A}}_c|}{r^*} \geq \epsilon_0\right)$ 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 $P\left(\frac{\sum_{j\in\mathcal{A}^*}(1-S_j)}{r^*} \geq (1-c)\epsilon_0\right) \geq P(\frac{|\overline{\mathcal{A}}_c|}{r^*} \geq \epsilon_0)$, 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 $\frac{|\mathcal{A}_c|}{r^*} \xrightarrow{p} 0$. □

A.3 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

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

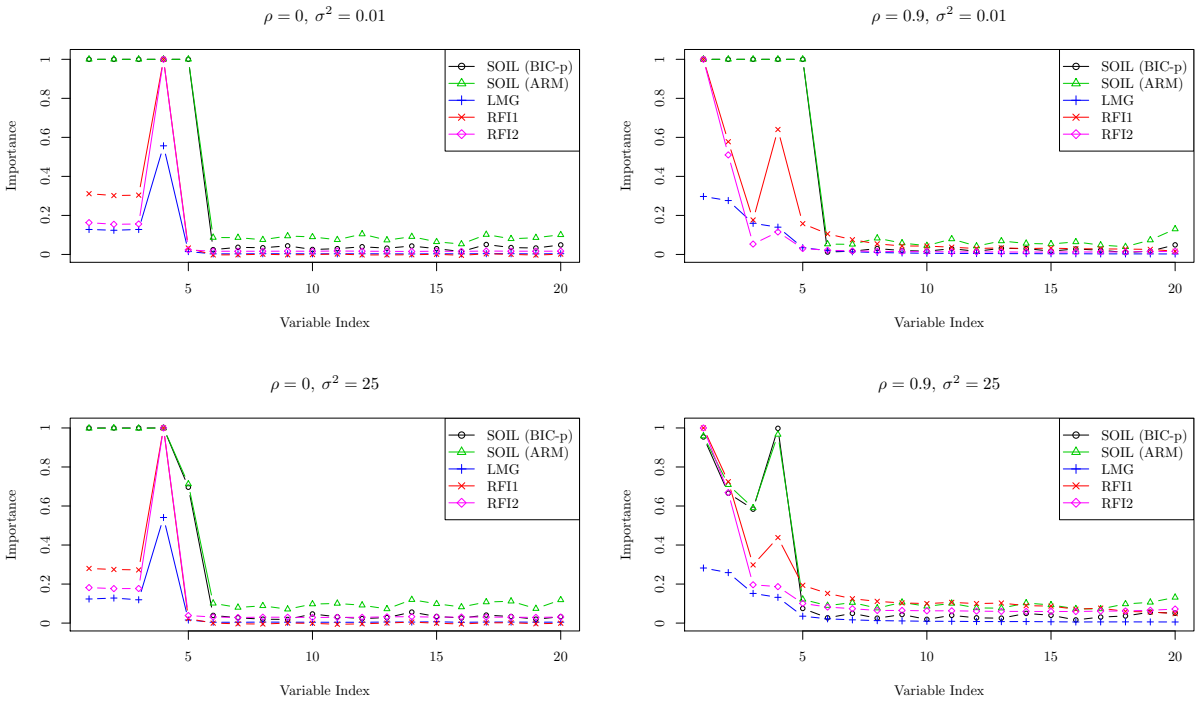
where $RSS_{\mathcal{A}^k}$ is the residual sum of squares of $\mathcal{A}^k$. For a practical reason, the authors approximate the above fiducial probability by

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

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.

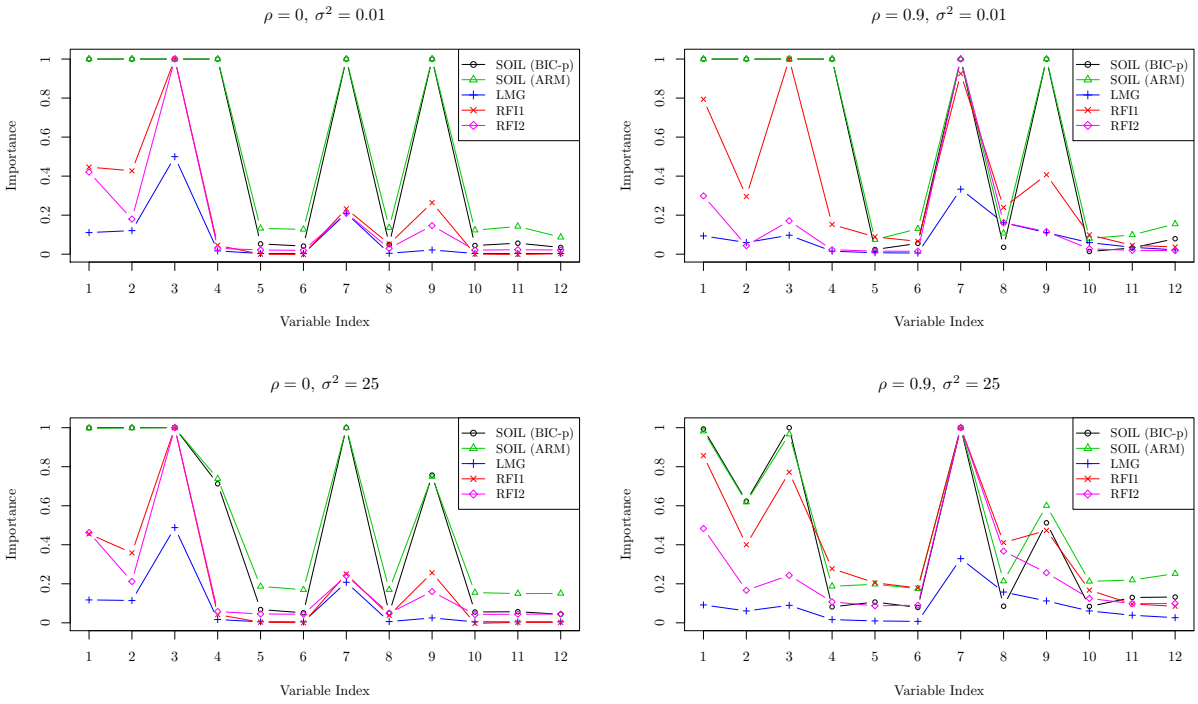
A.4 Additional simulation results

In this part, we provide the results of Example A.1- A.6, whose settings are described in Table 2.1 of the main body of the article. These results support our conclusions as discussed in Section 2.5.1.



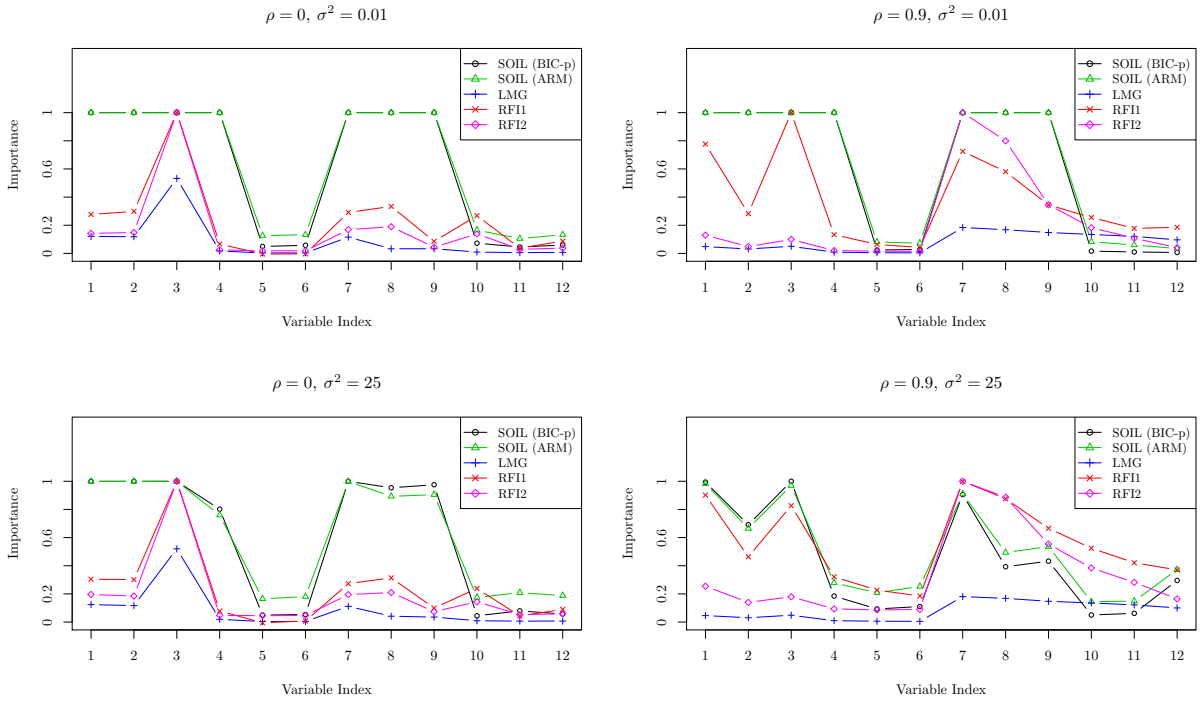
[H]

Figure A.1: Simulation results for Example A.1, where $n = 150$, $p = 20$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)$.



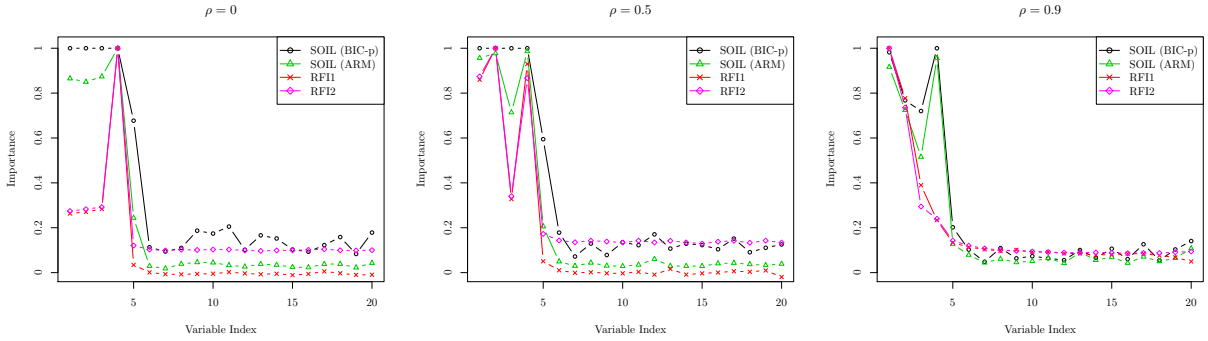
[H]

Figure A.2: Simulation results for Example A.2, 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 $(\beta_7^*, \beta_8^*, \ldots, \beta_{12}^*)^\intercal = (4, 0, 1, 0, 0, 0)^\intercal$.



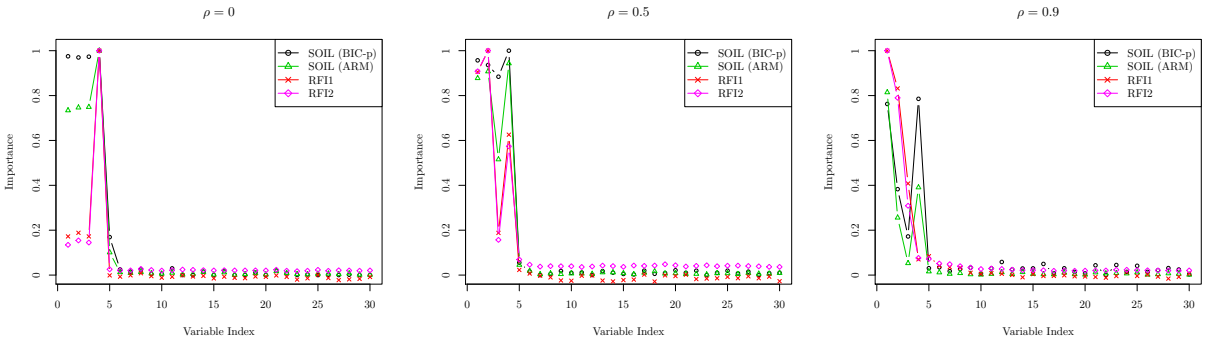
[H]

Figure A.3: Simulation results for Example A.3, 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 $(\beta_7^*, \beta_8^*, \ldots, \beta_{12}^*)^\intercal = (4, 2, 2, 0, 0, 0)^\intercal$.



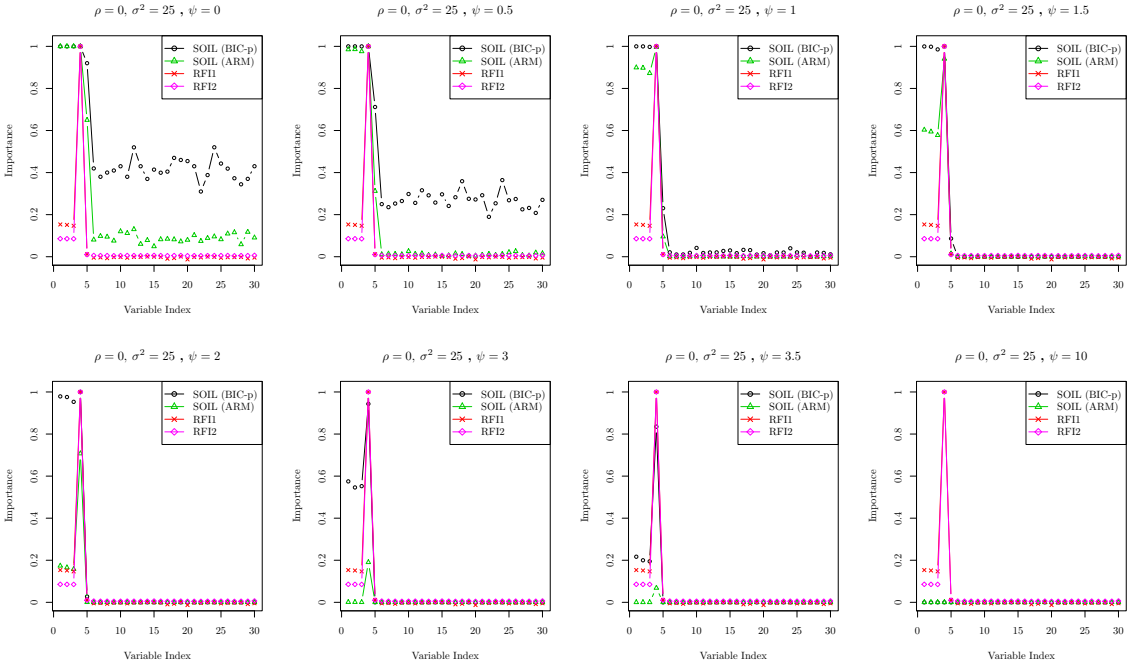
[H]

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



[H]

Figure A.5: Simulation results for Example A.5, where $n = 100$, $p = 200$. The true coefficients $\boldsymbol{\beta}^* = (4, 4, 4, -6\sqrt{2}, \frac{3}{4}, 0, ..., 0)$.



[H]

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

A.5 Comparison with stability selection

In this subsection, we present a comparison of SS (Meinshausen and 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 A.1.

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 A.1: Simulation settings for SS

It can be seen from Tables A.2 and A.3 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

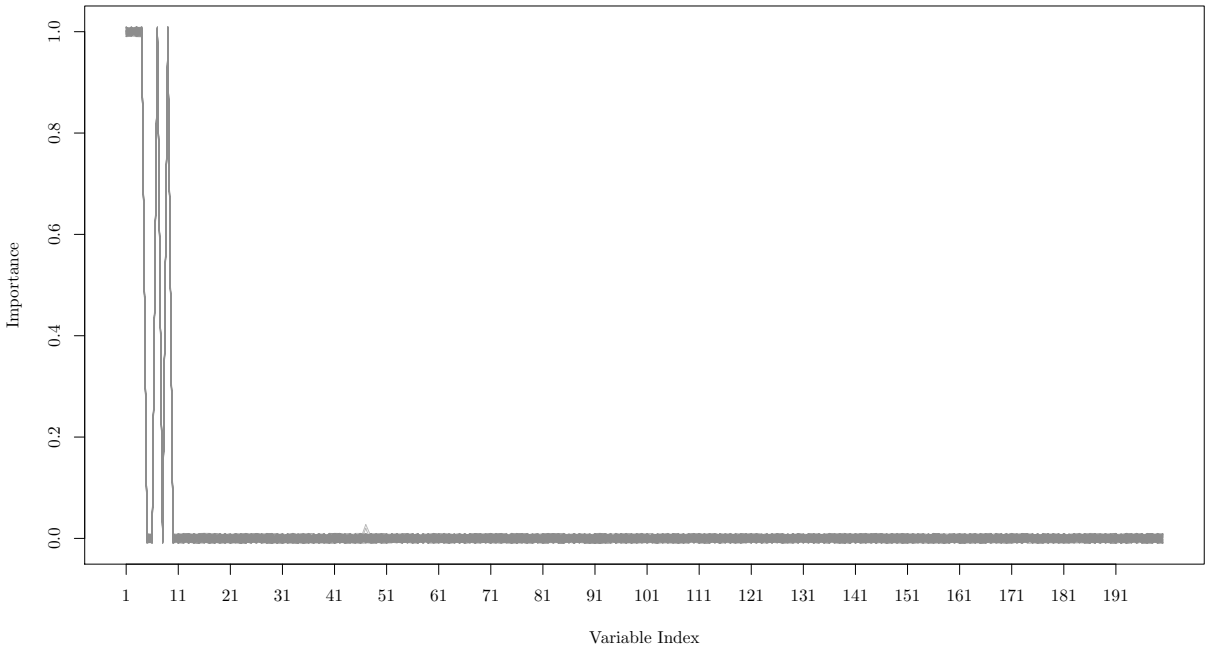
Table A.2: Variable importance for Example 1.

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 A.3: Variable importance for Example 2.

A.6 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 A.7. 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.



[H]

Figure A.7: 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.

Appendix B

Proofs and Figures of Chapter 3

B.1 Proof of Theorem 3

Proof B.1

Recall that

$$\mathrm{CVIL_p}(X^j;\delta,n,\pi_k) := \frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)})-Y_i)^2}{\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i)-Y_i)^2}-1$$

Denote the above numerator as A_n and the denominator as B_n. Let $A_{g,n} := \mathrm{E}_{\mathbf{X}_i,Y_i}(g_{\delta,n}(\mathbf{X}_i^{(j)})-Y_i)^2$ and $B_{g,n} := \mathrm{E}_{\mathbf{X}_i,Y_i}(g_{\delta,n}(\mathbf{X}_i)-Y_i)^2$. Let $A_g := \mathrm{E}_{\mathbf{X}_i,Y_i}(g_{\delta}(\mathbf{X}_i^{(j)})-Y_i)^2$ and $B_g := \mathrm{E}_{\mathbf{X}_i,Y_i}(g_{\delta}(\mathbf{X}_i)-Y_i)^2$. If we prove $A_n - A_{g,n_1} \xrightarrow{p} 0$ and $B_n - B_{g,n_1} \xrightarrow{p} 0$ as $n\to\infty$, it follows by Slutsky's theorem that

$$\begin{aligned}
\mathrm{CVIL_p}(X^j;\delta,n,\pi_k) - \mathrm{VI_p}(X_j;\delta,n) &= \frac{A_n}{B_n} - \frac{A_{g,n_1}}{B_{g,n_1}} \\
&= \frac{A_nB_{g,n_1} - A_{g,n_1}B_n}{B_nB_{g,n_1}} \\
&= \frac{(A_n - A_{g,n_1})B_{g,n_1} + A_{g,n_1}(B_{g,n_1}-B_n)}{B_nB_{g,n_1}} \\
&\xrightarrow{p} \frac{0}{B_g} + \frac{A_g}{B_g^2}0 = 0
\end{aligned}$$

as $n \to \infty$ for any $k = 1, ..., K$. Then we have

$$\frac{1}{K}\sum_{k=1}^{K} \mathrm{CVIL_p}(X^j; \delta, n, \pi_k) - \mathrm{VI_p}(X_j; \delta, n) \xrightarrow{p} 0$$

as $n \to \infty$. The desired result follows.

Thus it remains to prove $A_n - A_{g,n_1} \xrightarrow{p} 0$ and $B_n - B_{g,n_1} \xrightarrow{p} 0$ as $n \to \infty$. First,

$$\begin{aligned} A_n &= \frac{1}{n_2}\sum_{i=n_1+1}^{n} (\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - f(\mathbf{X}_i) - \epsilon_i)^2 \\ &:= \frac{1}{n_2}\sum_{i=n_1+1}^{n} A_{n_1 i}, \end{aligned}$$

For any constant $\varepsilon > 0$, we have

$$\begin{aligned} & P(|A_n - A_{g,n_1}| > \varepsilon) \\ \overset{(1)}{\leq} & \frac{1}{\varepsilon^2}\mathrm{E}\,|A_n - A_{g,n_1}|^2 \\ = & \frac{1}{\varepsilon^2 n_2^2}\mathrm{E}\left(\sum_{i=n_1+1}^{n}(A_{n_1 i} - A_{g,n_1})\right)^2 \\ = & \frac{1}{\varepsilon^2 n_2^2}\mathrm{E}_{\mathbf{Z}_1}\mathrm{E}\left[\left(\sum_{i=n_1+1}^{n}(A_{n_1 i} - A_{g,n_1})\right)^2 |\mathbf{Z}_1\right] \\ \overset{(2)}{=} & \frac{1}{\varepsilon^2 n_2^2}\mathrm{E}_{\mathbf{Z}_1}\left\{(n_2^2 - n_2)\mathrm{E}^2\left[(A_{n_1 i} - A_{g,n_1})|\mathbf{Z}_1\right] + n_2\mathrm{E}\left[(A_{n_1 i} - A_{g,n_1})^2|\mathbf{Z}_1\right]\right\} \quad \text{(B.1)} \\ = & \frac{1}{\varepsilon^2}\left\{(1 - \frac{1}{n_2})\underbrace{\mathrm{E}^2(A_{n_1 i} - A_{g,n_1})}_{(i)} + \underbrace{\frac{1}{n_2}\mathrm{E}(A_{n_1 i} - A_{g,n_1})^2}_{(ii)}\right\}, \end{aligned}$$

where (1) follows from Chebyshev's inequality and the conditional independency of the observations assures (2). It suffices to prove that both (i) and (ii) converge to 0

as $n \to \infty$.

$$\begin{aligned}
(i) &= \mathrm{E}^2\left[\left(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - f(\mathbf{X}_i) - \epsilon_i\right)^2 - A_{g,n_1}\right] \\
&= \mathrm{E}^2\left[\left(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - f(\mathbf{X}_i)\right)^2 - A_{g,n_1}\right] \\
&= \mathrm{E}^2\left[\hat{\delta}^2_{n_1}(\mathbf{X}_i^{(j)}) - g^2_{\delta,n_1}(\mathbf{X}_i^{(j)}) - 2f(\mathbf{X}_i)(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - g_{\delta,n_1}(\mathbf{X}_i^{(j)}))\right] \\
&\leq \mathrm{E}^2\left[|\hat{\delta}^2{}_{n_1}(\mathbf{X}_i^{(j)}) - g^2_{\delta,n_1}(\mathbf{X}_i^{(j)})| + 2||f||_\infty|\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - g_{\delta,n_1}(\mathbf{X}_i^{(j)})|\right] \\
&\overset{(3)}{\leq} 4(\max\{||f||_\infty, \sup_{n_1}||\hat{\delta}_{n_1}||, \sup_{n_1}||g_{\delta,n_1}||\})^2 \\
&\quad \cdot \mathrm{E}^2\left[\left|\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - g_{\delta,n_1}(\mathbf{X}_i^{(j)})\right|\right],
\end{aligned}$$

where (3) follows from the almost surely (uniform) boundedness of the functions f, $\hat{\delta}_{n_1}$ and g_{δ,n_1}. Since

$$\begin{aligned}
\mathrm{E}\left|\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - g_{\delta,n_1}(\mathbf{X}_i^{(j)})\right| &= \mathrm{E}_{\mathbf{Z}_1}\mathrm{E}\left[\left|\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - g_{\delta,n_1}(\mathbf{X}_i^{(j)})\right| |\mathbf{Z}_1\right] \\
&= \mathrm{E}_{\mathbf{Z}_1}||\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - g_{\delta,n_1}(\mathbf{X}_i^{(j)})||_1 \\
&\overset{(4)}{\leq} \mathrm{E}_{\mathbf{Z}_1}||\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - g_{\delta,n_1}(\mathbf{X}_i^{(j)})||_2 \\
&\overset{(5)}{\to} 0
\end{aligned}$$

where (4) follows from the monotonicity of the L_q norm and (5) follows from the boundedness of the functions and condition (A2). Thus (i) converges to 0 as n goes to infinity. By the (uniform) boundedness of the functions (conditions (A5) and (A6)), we have that $\mathrm{E}(A_{n_1 i} - A_{g,n_1})^2$ is bounded above regardless of n_1 and thus $(ii) = \frac{1}{n_2}\mathrm{E}(A_{n_1 i} - A_{g,n_1})^2 \to 0$ as $n \to \infty$. Hence $A_n - A_{g,n_1} \xrightarrow{p} 0$ as $n \to \infty$.

Following the same arguments, we have $B_n - B_{g,n_1} \to 0$ in probability as $n_1 \to \infty$. This completes the proof of Theorem 3. □

B.2 Proof of Theorem 4

Proof B.2

Following the same arguments as in proof of Theorem 3, with $\hat{\delta}_{n_1}^{(-j)}(\mathbf{X}_i^{(-j)})$, $g_{\delta,n_1}^{(-j)}(\mathbf{X}^{(-j)})$ replacing $\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)})$, $g_{\delta,n_1}(\mathbf{X}^{(j)})$ respectively, we can prove Theorem 4. □

B.3 Proof of Theorem 5

Proof B.3

Under one data splitting, the cross-validation based position variable importance for X^j is

$$\mathrm{CVIL_p}(X^j;\delta,n,\pi_k) \;=\; \frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)})-Y_i)^2}{\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i)-Y_i)^2}-1.$$

Conditional on the training data $\mathbf{Z_1}$, the 2-dimensional variables $\eta_i \triangleq \begin{pmatrix} (\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)})-Y_i)^2-\mu_{\delta^{(j)}} \\ (\hat{\delta}_{n_1}(\mathbf{X}_i)-Y_i)^2-\mu_{\delta} \end{pmatrix}$, $i = n_1,...,n$, are independent and identically distributed, with mean 0 and covariance matrix $\Sigma_n = \begin{pmatrix} \sigma^2_{\delta^{(j)}} & \sigma_{\delta,\delta^{(j)}} \\ \sigma_{\delta,\delta^{(j)}} & \sigma^2_{\delta} \end{pmatrix}$. By multivariate Berry-Esseen theorem, for any convex set $D\subseteq\mathbb{R}^2$, we have

$$|P(\frac{1}{\sqrt{n_2}}\sum_{i=n_1}^{n}\eta_i\in D|\mathbf{Z_1})-P(\xi\in D)|\le c\frac{2^{\frac{1}{4}}}{\sqrt{n_2}}\mathrm{E}[||\Sigma_n^{-\frac{1}{2}}\eta_i||^3|\mathbf{Z_1}],$$

where ξ is a 2-dimensional Gaussian with mean 0 and covariance matrix Σ_n, c is a universal constant and $||\cdot||$ is the Euclidean norm. Taking expectation $\mathrm{E}_{\mathbf{Z}_1}$ on both sides of the inequality, we have

$$\mathrm{E}_{\mathbf{Z}_1}|P(\frac{1}{\sqrt{n_2}}\sum_{i=n_1}^{n}\eta_i\in D|\mathbf{Z_1})-P(\xi\in D)|\le c\frac{2^{\frac{1}{4}}}{\sqrt{n_2}}\mathrm{E}_{\mathbf{Z}_1}\mathrm{E}[||\Sigma_n^{-\frac{1}{2}}\eta_i||^3|\mathbf{Z_1}],$$

which leads to

$$\begin{aligned}
|P(\frac{1}{\sqrt{n_2}}\sum_{i=n_1}^{n}\eta_i \in D) - P(\xi \in D)| &= |\mathrm{E}_{\mathbf{Z}_1} P(\frac{1}{\sqrt{n_2}}\sum_{i=n_1}^{n}\eta_i \in D|\mathbf{Z}_1) - P(\xi \in D)| \\
&\leq \mathrm{E}_{\mathbf{Z}_1}|P(\frac{1}{\sqrt{n_2}}\sum_{i=n_1}^{n}\eta_i \in D|\mathbf{Z}_1) - P(\xi \in D)| \\
&\leq c\frac{2^{\frac{1}{4}}}{\sqrt{n_2}}\mathrm{E}_{\mathbf{Z}_1}\mathrm{E}[||\Sigma_n^{-\frac{1}{2}}\eta_i||^3|\mathbf{Z}_1] \\
&= c\frac{2^{\frac{1}{4}}}{\sqrt{n_2}}\mathrm{E}||\Sigma_n^{-\frac{1}{2}}\eta_i||^3.
\end{aligned}$$

By condition (B4), we have the right hand side converges to 0 in probability as n goes to infinity. Thus,

$$\frac{1}{\sqrt{n_2}}\Sigma_n^{-1/2}\sum_{i=n_1}^{n}\eta_i \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, I_2\right)$$

as $n \to \infty$.

Due to the boundedness and the convergence of $g_{\delta,n}$ to g_δ, it is not hard to prove $\sigma_{\delta^{(j)}} \xrightarrow{p} \sigma^0_{g^{(j)}}$, $\sigma_{\delta,\delta^{(j)}} \xrightarrow{p} \sigma^0_{g,g^{(j)}}$, $\sigma_{\delta^{(j)}} \xrightarrow{p} \sigma^0_{g^{(j)}}$, $\mu_{\delta^{(j)}} \xrightarrow{p} \mu^0_{g^{(j)}}$ and $\mu_\delta \xrightarrow{p} \mu^0_g$ as $n \to \infty$. Together with $\sqrt{n_2}\cdot||\hat{\delta}_{n_1}(\mathbf{x}) - g_\delta(\mathbf{x})||_1 \xrightarrow{p} 0$ and $\sqrt{n_2}\cdot||\hat{\delta}_{n_1}(\mathbf{x}^{(j)}) - g^{(j)}_\delta(\mathbf{x}^{(j)})||_1 \xrightarrow{p} 0$, by continuous mapping theorem, we have

$$\sqrt{n_2}\begin{pmatrix}(\sigma^0_{g^{(j)}})^2 & \sigma^0_{g,g^{(j)}}\\ \sigma^0_{g,g^{(j)}} & (\sigma^0_g)^2\end{pmatrix}^{-1/2}\left(\begin{pmatrix}\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - Y_i)^2\\ \frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i) - Y_i)^2\end{pmatrix} - \begin{pmatrix}\mu^0_{g^{(j)}}\\ \mu^0_g\end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, I_2\right).$$

as $n \to \infty$. By delta method, we have

$$\sqrt{n_2}(\frac{\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i^{(j)}) - Y_i)^2}{\frac{1}{n_2}\sum_{i=n_1+1}^{n}(\hat{\delta}_{n_1}(\mathbf{X}_i) - Y_i)^2} - \frac{\mu^0_{g^{(j)}}}{\mu^0_g}) \xrightarrow{d} N\left(0, \left(\frac{\sigma^0_{g^{(j)}}}{\mu^0_g}\right)^2 + \left(\frac{\mu^0_{g^{(j)}}\sigma^0_g}{(\mu^0_g)^2}\right)^2 - 2\frac{\mu^0_{g^{(j)}}\sigma^0_{g,g^{(j)}}}{(\mu^0_g)^3}\right)$$

as $n \to \infty$, which completes the proof. □

B.4 Figures

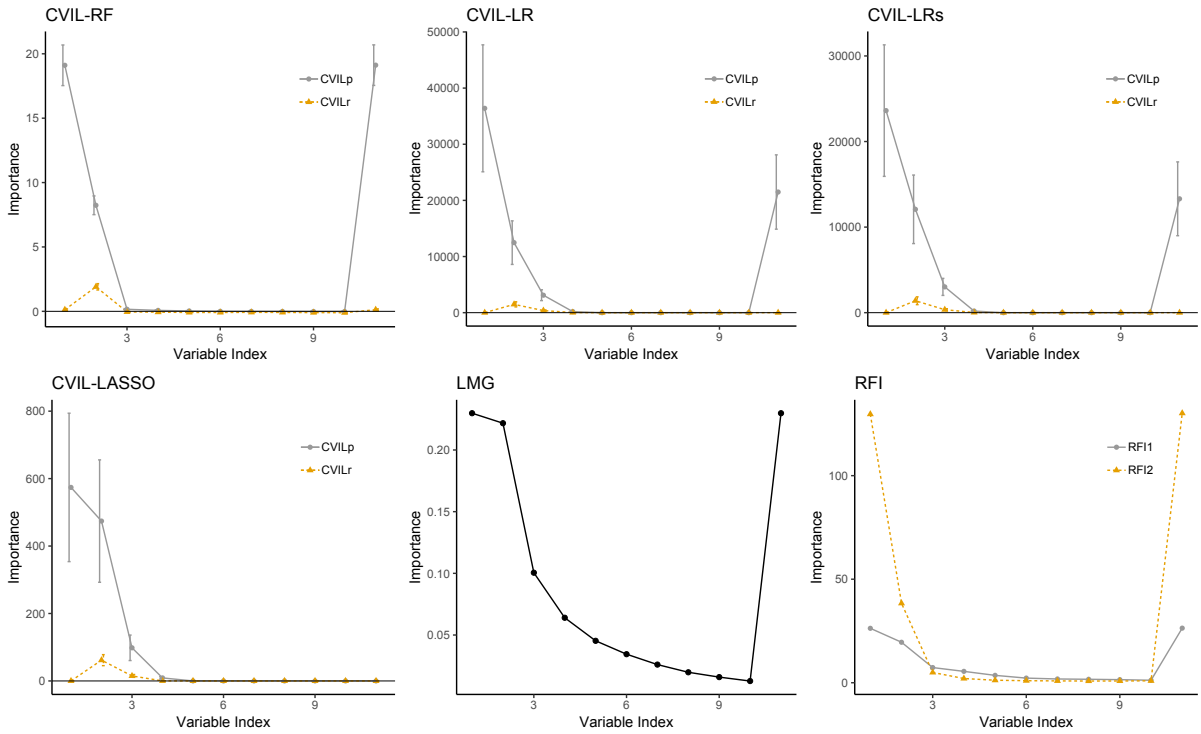


Figure B.1: Example 1, $\rho = 0.9$

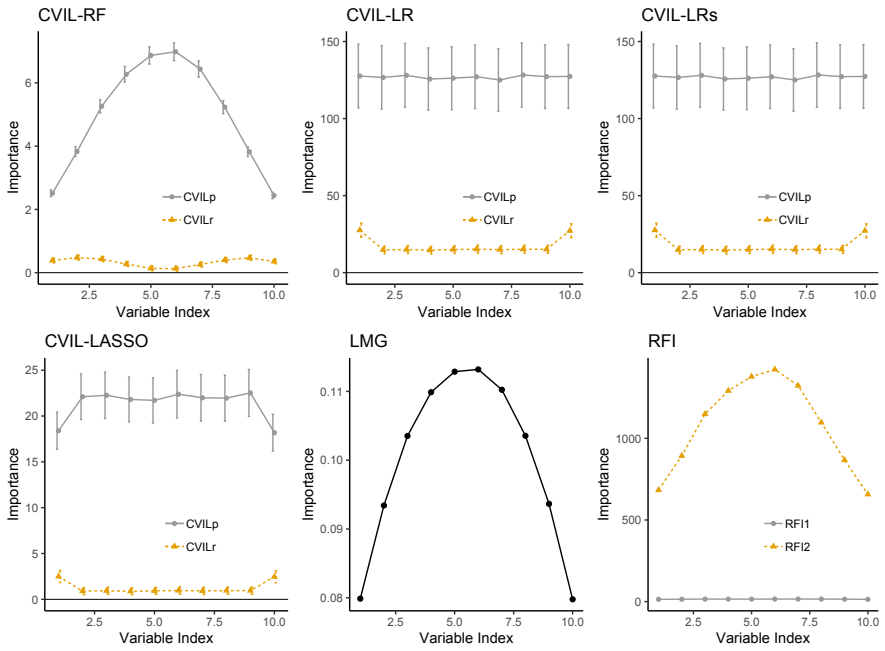


Figure B.2: Example 2, $c = 1$, $\rho = 0.9$

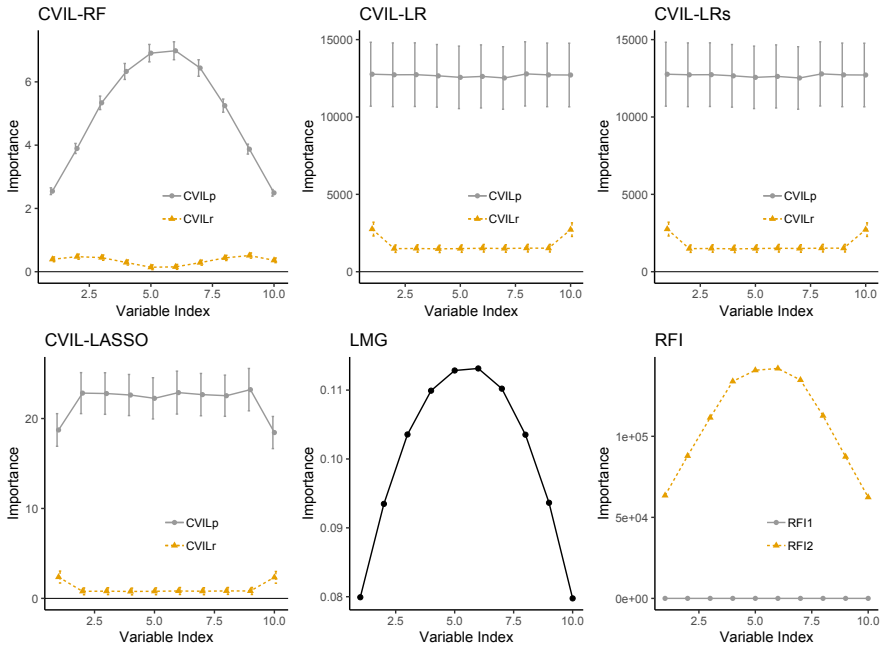


Figure B.3: Example 2, $c = 10$, $\rho = 0.9$

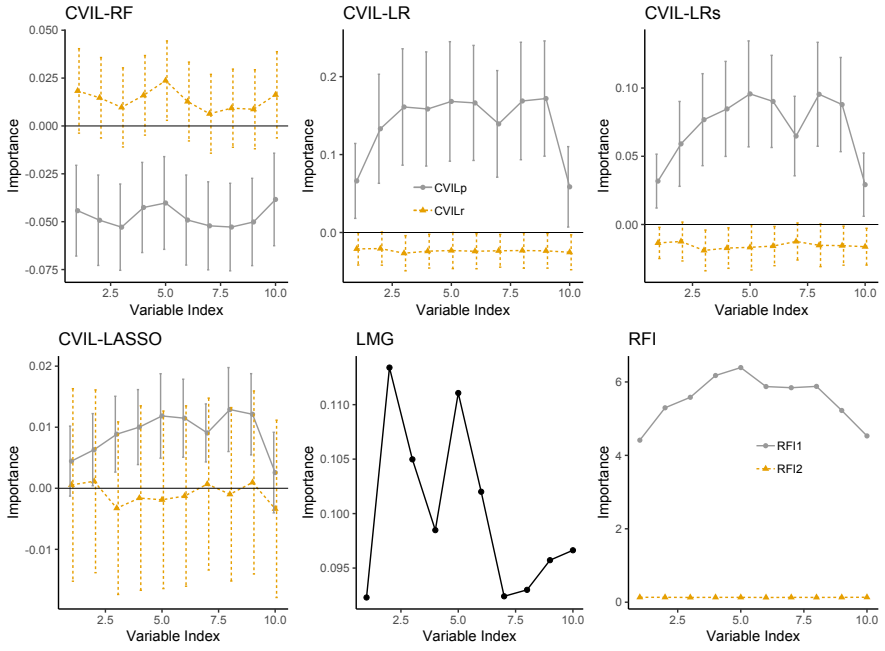


Figure B.4: Example 2, $c = 0$, $\rho = 0.9$

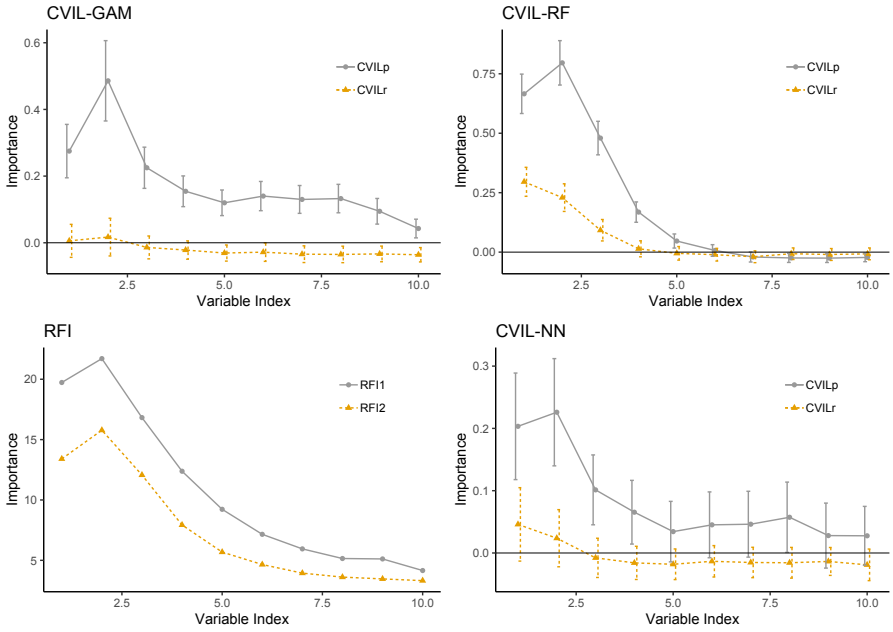


Figure B.5: Example 3, $\rho = 0.9$

Appendix C

Proofs of Chapter 4

C.1 Proof of Theorem 7

C.1.1 Proof of the Upper Bound ((4.4))

Recall that $h(\mathbf{z}) = \mathbf{z}^T\beta$ and $\hat{h}(\mathbf{z}) = \mathbf{z}^T\hat{\beta}$. Set $\mathbf{h}_{\mathtt{I}} := P_{\mathtt{I}}\mathbf{h}$ as the estimator by model $\mathtt{I}$, where we use the bold-face $\mathbf{h} = (h(\mathbf{z}_1^T), ..., h(\mathbf{z}_n^T))^T$ to denote the mean regression function vector and $\mathbf{z}_i$ is the i-th row of the full design matrix $\mathbf{Z}$. We first prove that $\hat{\mathtt{I}}$ is equivalently an ABC estimator over the candidate set we consider. The SRC assumption with $l_1 = r_1$, $l_2 = r_2$ assures that $r_1 + r_2 \leq n$. It follows that, for any model $\mathtt{I} = (\mathtt{I}_1, \mathtt{I}_2)$ with $|\mathtt{I}_1|_0 = r_1$, $|\mathtt{I}_2|_0 = r_2$, the corresponding submatrix $\mathbf{Z}_{\mathtt{I}}$ is full rank, i.e., $r_{\mathtt{I}} = r_1 + r_2$. Thus,

$$\begin{aligned}
\hat{\mathtt{I}} &= \arg\min_{\mathtt{I}\in\mathcal{F}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{\mathtt{I}})^2 \\
&= \arg\min_{\mathtt{I}\in\mathcal{F}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{\mathtt{I}})^2 + 2r_{\mathtt{I}}\sigma^2 + \lambda\sigma^2 C_{\mathtt{I}} \\
&= \arg\min_{\mathtt{I}\in\mathcal{F}} ABC(\mathtt{I}),
\end{aligned}$$

where $\mathcal{F}$ is the collection of models that have r_1 non-zero main effects and r_2 non-zero interaction effects with $0 \leq r_1 \leq p_n, 1 \leq r_2 \leq \binom{r_1}{2}$, and all the models in $\mathcal{F}$ share the

same model descriptive complexity

$$C_{\mathtt{I}^{strong}_{r_1,r_2}} = \log \binom{p_n}{r_1} + \log \binom{\binom{r_1}{2}}{r_2}.$$

The ABC criterion and the model descriptive complexity are introduced near ((4.10)). Therefore, $\hat{\mathtt{I}}$ is an ABC estimator over the candidate set $\mathcal{F}$.

Next we prove the upper bound. Since $\hat{\mathtt{I}}$ is an ABC estimator over the candidate set $\mathcal{F}$, by Theorem 1 in Yang (1999), we have:

$$E(\mathcal{L}(\hat{\mathtt{I}})) \leq c \inf_{\mathtt{I}\in\mathcal{F}} \left(\frac{1}{n} \|\mathbf{h}_{\mathtt{I}} - \mathbf{h}\|_2^2 + \frac{\sigma^2 r_{\mathtt{I}}}{n} + \frac{\lambda\sigma^2 C_{\mathtt{I}}}{n} \right), \tag{C.1}$$

where c is a positive constant that depends on the constant λ only. When $h \in \mathcal{W} = \mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})$, there exists a specific model in $\mathcal{F}$ such that the projection estimator of this model is equal to $\mathbf{h}$. We consider the RHS of ((C.1)) evaluated at such a model, where we still denote it as $\mathtt{I}_{r_1,r_2}$ for convenience. Thus,

$$\begin{aligned} E(\mathcal{L}(\hat{\mathtt{I}})) &\leq c\left(\left\|\mathbf{h}_{\mathtt{I}_{r_1,r_2}} - \mathbf{h}\right\|^2 + \frac{\sigma^2 r_{\mathtt{I}_{r_1,r_2}}}{n} + \frac{\lambda\sigma^2 C_{\mathtt{I}_{r_1,r_2}}}{n} \right) \\ &= \underbrace{\frac{c}{n}\left(\sigma^2 r_{\mathtt{I}_{r_1,r_2}} + \lambda\sigma^2 C_{\mathtt{I}_{r_1,r_2}}\right)}_{(i)}. \end{aligned}$$

The term (i) is bounded as follows:

$$\begin{aligned} (i) &\leq \frac{c_1\lambda}{n}\sigma^2 \left(\frac{1}{\lambda}(r_1 + r_2) + \log \binom{p_n}{r_1} + \log \binom{\binom{r_1}{2}}{r_2} \right) \\ &\leq \frac{c_1\lambda}{n}\sigma^2 \left(\frac{1}{\lambda}(r_1 + r_2) + r_1\left(1 + \log\frac{p_n}{r_1}\right) + r_2\left(1 + \log\frac{\binom{r_1}{2}}{r_2}\right) \right) \\ &\leq \frac{c_2}{n}\sigma^2 \left(r_1\left(1 + \log\frac{p_n}{r_1}\right) + r_2\left(1 + \log\frac{\binom{r_1}{2}}{r_2}\right) \right). \end{aligned}$$

Therefore,

$$E(\mathcal{L}(\hat{\mathtt{I}})) \leq \frac{c_2 \cdot \sigma^2}{n}\left(r_1\left(1+\log\frac{p_n}{r_1}\right)+r_2\left(1+\log\frac{\binom{r_1}{2}}{r_2}\right)\right).$$

Thus we have

$$\min_{\hat{h}}\max_{h\in\mathcal{W}} E\mathsf{L}(\hat{h},h) \leq \max_{h\in\mathcal{W}} E(\mathcal{L}(\hat{\mathtt{I}})) \leq \frac{c_2 \cdot \sigma^2}{n}\left(r_1\left(1+\log\frac{p_n}{r_1}\right)+r_2\left(1+\log\frac{\binom{r_1}{2}}{r_2}\right)\right),$$

where the above c_1, c_2 are universal constants.

C.1.2 Proof of the Lower Bound ((4.5))

Before stating the proof of ((4.5)), we introduce the local metric entropy, two important sets that aid the understanding of the metric entropy of the regression function space, together with the lemmas in relation to these two sets.

Metric Entropy

Metric entropy plays a central role in minimax theory, through the concepts of packing and covering. It provides a way to understand the "cardinality" of a set with infinitely many elements. In deriving the lower bound, information theoretic techniques play a key role, such as the local metric entropy, Fano's inequality, Shannon's mutual information and Kullback–Leibler divergence. We begin by introducing the definition of the local metric entropy.

Definition 5 (Local Metric Entropy) Given a metric space $(\mathcal{X},\rho)$, let $B(x,\epsilon) = \{x' \in \mathcal{X} | \rho(x,x') \leq \epsilon\}$ be a ϵ-ball around x. For $0 < a < 1$, the **a-local ϵ-entropy at** x, denoted as $\log M_x^a(\epsilon;\mathcal{X},\rho)$, is defined as the $a\epsilon$-packing entropy of $B(x,\epsilon)$. The **a-local ϵ-entropy**, denoted as $\log M_{\mathtt{local}}^a(\epsilon;\mathcal{X},\rho)$, is then defined as the maximum (or supremum if maximum does not exist) of $\log M_x^a(\epsilon;\mathcal{X},\rho)$ over all x in $\mathcal{X}$, i.e., $\log M_{\mathtt{local}}^a(\epsilon;\mathcal{X},\rho) = \max_{x\in\mathcal{X}} \log M_x^a(\epsilon;\mathcal{X},\rho)$. □

Important Subsets

Set the Hamming distance between any two vectors $v, v' \in \mathbb{R}^d$ as $\rho_H(v, v') = \sum_{i=1}^{d} \mathbb{1}_{v_i \neq v_i'}$. Consider the set

$$\mathcal{H} = \left\{ \beta \in \ddot{\mathbb{R}}^{p_n}_{strong} : \beta \in \{-1, 0, 1\}^{p_n + \binom{p_n}{2}}, \left\|\beta^{(1)}\right\|_0 \leq r_1, \left\|\beta^{(2)}\right\|_0 \leq r_2 \right\}$$

and let $\mathcal{H}_1$ denote a subset of $\mathcal{H}$ where the the first r_1 coordinates are fixed, i.e.,

$$\mathcal{H}_1 = \left\{ \beta \in \mathcal{H} : \beta^{(1)} = (\underbrace{1, ..., 1}_{r_1}, \underbrace{0, ..., 0}_{p_n - r_1}), \left\|\beta^{(2)}\right\|_0 = r_2 \right\}.$$

Let $\mathcal{H}_2$ denote another subset of $\mathcal{H}$ where no interaction effect exists, i.e.,

$$\mathcal{H}_2 = \left\{ \beta \in \mathcal{H} : \left\|\beta^{(1)}\right\|_0 = r_1, \left\|\beta^{(2)}\right\|_0 = 0 \right\},$$

The following two lemmas of the metric entropy of the subsets $\mathcal{H}_1$ and $\mathcal{H}_2$ are needed in the proof of ((4.5)).

Lemma 1 If $r_2 \leq \frac{2}{3}\binom{r_1}{2}$, then there exists a subset of $\mathcal{H}_1$ with its cardinality no less than $\exp\left(\frac{r_2}{2} \log \frac{\binom{r_1}{2} - r_2/2}{r_2}\right)$ such that the pairwise Hamming distance of the points in this subset is greater than $r_2/2$. □

Proof C.1

The proof is presented in Appendix C.1.3. □

Lemma 2 If $r_1 \leq 2p_n/3$, then there exists a subset of $\mathcal{H}_2$ with its cardinality no less than $\exp\left(\frac{r_1}{2} \log \frac{p_n - r_1/2}{r_1}\right)$ such that the pairwise Hamming distance of the points in this subset is greater than $r_1/2$. □

Proof C.2

The proof is similar to that of Lemma 1. □

Proof of ((4.5))

It suffices to prove under $r_2 \leq (r_1^2 - r_1)/4$. Since $r_2(1 + \log(\binom{r_1}{2}/r_2)) \asymp \binom{r_1}{2}$ for $\frac{1}{2}\binom{r_1}{2} \leq r_2 \leq \binom{r_1}{2}$, the monotonicity of the minimax risk in the function class reduces the proof to the case $r_2 \leq (r_1^2 - r_1)/4$. Similarly it suffices to prove under $r_1 \leq p_n/2$.

Recall that $B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong}) = \left\{\beta \in \ddot{\mathbb{R}}^{p_n}_{strong} : \left\|\beta^{(1)}\right\|_0 \leq r_1, \left\|\beta^{(2)}\right\|_0 \leq r_2\right\}$ is the coefficient space of interest and $\mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong}) = \left\{h : h(\mathbf{z}) = \mathbf{z}^T\beta, \beta \in B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})\right\}$ is the mean regression function space. For convenience, let h_θ, h_ϑ denote the regression functions with coefficents θ, ϑ respectively, i.e., $h_\theta(\mathbf{z}) = \mathbf{z}^T\theta$, $h_\vartheta(\mathbf{z}) = \mathbf{z}^T\vartheta$. Let

$$B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(\epsilon) = \left\{\beta : \beta \in B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong}), \|\beta\|_2 \leq \epsilon\right\}$$

be an l_2-ball of radius ϵ around 0 in $B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})$ and

$$\mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(h, \epsilon_0) = \left\{h' : h'(\mathbf{z}) = \mathbf{z}^T\beta, \beta \in B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong}), d(h', h) \leq \epsilon_0\right\}$$

be the ball of radius ϵ_0 around the underlying regression function h. Without loss of generality, we assume $h = 0$. The square root of the empirical l_2-norm loss $d(h_\theta, h_\vartheta) := \sqrt{\frac{1}{n}\sum_{i=1}^n (h_\theta(\mathbf{z}_i) - h_\vartheta(\mathbf{z}_i))^2} = \frac{1}{\sqrt{n}}\|\mathbf{Z}(\theta - \vartheta)\|_2$ is used to measure the distance between any two functions h_θ, h_ϑ. We prove the following two cases separately.

Case 1: $\frac{r_1}{2}\log((p_n - r_1/2)/r_1) \leq \frac{r_2}{2}\log((\binom{r_1}{2} - r_2/2)/r_2)$. We consider the subset $\mathcal{H}_1' = \{\boldsymbol{\epsilon} \circ \beta : \beta \in \mathcal{H}_1\}$ of the l_2-ball $B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(\epsilon)$, where $\circ$ is the point-wise product of two vectors,

$$\boldsymbol{\epsilon} = \frac{\epsilon}{\sqrt{2}}(\underbrace{1/\sqrt{r_1}, ..., 1/\sqrt{r_1}}_{p_n}, \underbrace{1/\sqrt{r_2}, ..., 1/\sqrt{r_2}}_{(p_n^2 - p_n)/2})$$

and

$$\mathcal{H}_1 = \left\{\beta \in \mathcal{H} : \beta^{(1)} = (\underbrace{1, ..., 1}_{r_1}, \underbrace{0, ..., 0}_{p_n - r_1}), \left\|\beta^{(2)}\right\|_0 = r_2\right\}.$$

From Lemma 1, there exists a subset $\mathcal{H}_{sub}$ of $\mathcal{H}_1$ such that $|\mathcal{H}_{sub}| \geq \exp(\frac{r_2}{2}\log\frac{\binom{r_1}{2}-r_2/2}{r_2})$ and the pairwise Hamming distance of the elements within $\mathcal{H}_{sub}$ is greater than $r_2/2$. Set $\mathcal{H}'_{sub} := \{\boldsymbol{\epsilon}\circ\beta : \beta \in \mathcal{H}_{sub}\}$. For any $\theta', \vartheta' \in \mathcal{H}'_{sub}$, there exist $\theta, \vartheta \in \mathcal{H}_{sub}$ such that $\|\theta' - \vartheta'\|_2 = \|\boldsymbol{\epsilon}\circ\theta - \boldsymbol{\epsilon}\circ\vartheta\|_2 \geq \frac{\epsilon}{\sqrt{2r_2}}\sqrt{\rho_H(\theta,\vartheta)} \geq \frac{\epsilon}{\sqrt{2r_2}}\sqrt{r_2/2} = \frac{\epsilon}{2}$. We also have $|\mathcal{H}'_{sub}| = |\mathcal{H}_{sub}|$ since it is a one-to-one mapping from $\mathcal{H}_{sub}$ to $\mathcal{H}'_{sub}$. Thus, we have $\mathcal{H}'_{sub} \subseteq B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(\epsilon)$ and the pairwise l_2-distance of the elements in $\mathcal{H}'_{sub}$ is greater than $\epsilon/2$.

For any $\theta', \vartheta' \in \mathcal{H}'_{sub} \subseteq B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(\epsilon)$, let $h_{\theta'}, h_{\vartheta'}$ be such that $h_{\theta'}(\mathbf{z}) = \mathbf{z}^T\theta', h_{\vartheta'}(\mathbf{z}) = \mathbf{z}^T\vartheta'$. By SRC assumption with $l_1 = r_1$, $l_2 = r_2$, we have

$$b_1\frac{\epsilon}{2} \leq b_1\|(\theta' - \vartheta')\|_2 \leq d(h_{\theta'}, h_{\vartheta'})$$

$$d(h, h_{\vartheta'}) \leq b_2\|(0 - \vartheta')\|_2 \leq b_2\epsilon.$$

Let $\epsilon_0 = b_2\epsilon$, it follows that $\mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(h, \epsilon_0)$ has a subset

$$\mathcal{F}_{sub} := \left\{h' : h'(\mathbf{z}) = \mathbf{z}^T\beta, \beta \in \mathcal{H}'_{sub}, d(h', h) \leq \epsilon_0\right\},$$

in which the pairwise distance (in terms of d) of the functions are no less than $\frac{b_1}{2b_2}\epsilon_0$. This implies that the $\frac{b_1}{2b_2}$-local ϵ_0-packing entropy of $\mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(h, \epsilon_0)$ is lower bounded by $\log|\mathcal{F}_{sub}| = \log|\mathcal{H}'_{sub}| \geq \frac{r_2}{2}\log\frac{\binom{r_1}{2}-r_2/2}{r_2}$. So $\log M^{b_1/(2b_2)}_{\texttt{local}}(\epsilon_0)$ of $\mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})$ is no less than $\frac{r_2}{2}\log((r_1^2 - r_1 - r_2)/2r_2)$. Then by (7) in Yang and Barron (1999), the minimax risk is lower bounded by

$$c_1\frac{\sigma^2\frac{r_2}{2}\log(\frac{r_1^2-r_1-r_2}{2r_2})}{n} = c_1\frac{\sigma^2}{n}\left(\frac{r_1}{2}\log\frac{p_n - r_1/2}{r_1} \vee \frac{r_2}{2}\log\frac{\binom{r_1}{2} - r_2/2}{r_2}\right),$$

where $c_1 > 0$ is a constant that depends on b_1 and b_2 only.

Case 2: $\frac{r_1}{2}\log((p_n - r_1/2)/r_1) \geq \frac{r_2}{2}\log((\binom{r_1}{2} - r_2/2)/r_2)$. We consider the subset $\mathcal{H}'_2 = \epsilon'_1\mathcal{H}_2$ of $B_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})(\epsilon)$, where $\epsilon'_1 = \epsilon/\sqrt{r_1}$ and

$$\mathcal{H}_2 := \left\{\beta \in \mathcal{H} : \left\|\beta^{(1)}\right\|_0 = r_1, \left\|\beta^{(2)}\right\|_0 = 0\right\}.$$

Following the same arguments above, we conclude that the minimax is lower bounded by

$$c_2\frac{\sigma^2}{n}\frac{r_1}{2}\log\frac{p_n - r_1/2}{r_1} = c_2\frac{\sigma^2}{n}\left(\frac{r_1}{2}\log\frac{p_n - r_1/2}{r_1} \vee \frac{r_2}{2}\log\frac{\binom{r_1}{2} - r_2/2}{r_2}\right),$$

where $c_2 > 0$ is a constant that depends on b_1 and b_2 only.

Notice that when $p_n/r_1 \geq 2$, we have $\log(p_n/r_1 - \frac{1}{2}) \geq \frac{1}{10}(1+\log(p_n/r_1))$. Similarly, we have $\log\left(\binom{r_1}{2}/r_2 - \frac{1}{2}\right) \geq \frac{1}{10}\left(1+\log\left(\binom{r_1}{2}/r_2\right)\right)$ when $\binom{r_1}{2}/r_2 \geq 2$. Together with the fact that the lower bounds for the two cases are the same, the minimax risk is lower bounded by

$$c\frac{\sigma^2}{n}\left(r_1(1+\log(\frac{p_n}{r_1})) \vee r_2(1+\log\frac{\binom{r_1}{2}}{r_2})\right).$$

Thus the desired lower bound holds.

C.1.3 Proof of Lemma 1

First we have $|\mathcal{H}_1| = \binom{(r_1^2 - r_1)/2}{r_2}2^{r_2}$ since the main effects are fixed. Fix $z \in \mathcal{H}_1$, let $\mathcal{A}$ denote the collection of all the points in $\mathcal{H}_1$ that are within $\frac{r_2}{2}$ Hamming distances to z, i.e., $\mathcal{A} = \{z' \in \mathcal{H}_1 : \rho_H(z, z') \leq r_2/2\}$. It follows that the cardinality of $\mathcal{A}$ is bounded above:

$$|\mathcal{A}| \;\leq\; \binom{\binom{r_1}{2}}{r_2/2}3^{r_2/2}.$$

For this upper bound, since the main effects are fixed for any point in $\mathcal{H}_1$, we only need to pick $r_2/2$ positions of the interaction effects where z' is different from z. In the remaining interaction effect positions, z' is the same as z. It gives us at most $\binom{\binom{r_1}{2}}{r_2/2}$ possible choices of the $r_2/2$ positions out of the $\binom{r_1}{2}$ coordinates. For these $r_2/2$ positions, z' can take any values in $\{-1, 1, 0\}$, thus the desired upper bound follows.

Let $\mathcal{B}$ be a subset of $\mathcal{H}_1$ such that $|\mathcal{B}| \leq m := \binom{\binom{r_1}{2}}{r_2}/\binom{\binom{r_1}{2}}{r_2/2}$. Consider the collection of the points in $\mathcal{H}_1$ that are within $r_2/2$ Hamming distance to some element in $\mathcal{B}$,

i.e., $\{z \in \mathcal{H}_1 : \rho_H(z, z') \leq \frac{r_2}{2} \ for \ some \ z' \in \mathcal{B}\}$. We have

$$\begin{aligned} & \left|\left\{z \in \mathcal{H}_1 : \rho_H(z, z') \leq \frac{r_2}{2} \ for \ some \ z' \in \mathcal{B}\right\}\right| \\ \leq & \ |\mathcal{B}|\,|\mathcal{A}| \\ \leq & \ \frac{\binom{\binom{r_1}{2}}{r_2}}{\binom{\binom{r_1}{2}}{r_2/2}} \cdot \binom{\binom{r_1}{2}}{r_2/2} 3^{r_2/2} \\ < & \ \binom{\binom{r_1}{2}}{r_2} 2^{r_2} \\ = & \ |\mathcal{H}_1|\,. \end{aligned}$$

The strictly less inequality implies that for any set $\mathcal{B} \subset \mathcal{H}_1$ with $|\mathcal{B}| \leq m$, $\exists z \in \mathcal{H}_1$ such that $\rho_H(z, z') > \frac{1}{2}r_2$ for all $z' \in \mathcal{B}$. By induction, we can create a set $\mathcal{B} \subset \mathcal{H}_1$ with $|\mathcal{B}| > m$ such that Hamming distance between any two elements in $\mathcal{B}$ exceeds $\frac{1}{2}r_2$. Next, we introduce one useful inequality. When $0 \leq B \leq \frac{2}{3}A$ for $A, B \in \mathbb{N}$, we have

$$\frac{\binom{A}{B}}{\binom{A}{\frac{B}{2}}} = \frac{(A-\frac{B}{2})!(\frac{B}{2})!}{(A-B)!(B)!} = \prod_{j=1}^{B/2} \frac{A-B+j}{\frac{B}{2}+j} \geq \prod_{j=1}^{B/2} \frac{A-B+\frac{B}{2}}{\frac{B}{2}+\frac{B}{2}} = (\frac{A-\frac{B}{2}}{B})^{B/2}.$$

When $r_2 \leq (r_1^2 - r_1)/3$, we have

$$m \ = \ \frac{\binom{\binom{r_1}{2}}{r_2}}{\binom{\binom{r_1}{2}}{r_2/2}} \geq \left(\frac{\binom{r_1}{2} - r_2/2}{r_2}\right)^{r_2/2}.$$

Thus,

$$\log m \ \geq \ \frac{r_2}{2} \log \frac{\binom{r_1}{2} - \frac{r_2}{2}}{r_2}.$$

The desired result follows.

C.2 Proof of Theorem 8

Proof C.3

The proofs are similar to the arguments for strong heredity with slight differences.

To prove the upper bound under weak heredity, we instead consider the model $\hat{\mathtt{I}} = \arg\min_{\mathtt{I}\in\mathtt{I}^{weak}_{r_1,r_2}} \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{\mathtt{I}})^2$ that minimizes the residual sum of squares over all the models that have r_1 non-zero main effects and r_2 non-zero interaction effects under weak heredity. The model descriptive complexity is thus different from the strong heredity. In this case, $C_{\mathtt{I}^{weak}_{r_1,r_2}} = \log\binom{p_n}{r_1} + \log\binom{K}{r_2}$ with $K = r_1(p_n - (r_1+1)/2)$ for $1 \le r_1 \le p_n \wedge n$ and $0 \le r_2 \le (r_1 p_n - \binom{r_1}{2} - r_1) \wedge n$. The ABC criteria for the models are defined as in ((4.10)). The same arguments in the proof of ((4.4)) can then be used.

To prove the lower bound under weak heredity, we consider the set

$$\mathcal{H}_{weak} = \left\{\beta \in \ddot{\mathbb{R}}^{p_n}_{weak} : \beta \in \{-1,0,1\}^{p_n+\binom{p_n}{2}}, \left\|\beta^{(1)}\right\|_0 \le r_1, \left\|\beta^{(2)}\right\|_0 \le r_2\right\}.$$

Then the two important subsets are instead

$$\mathcal{H}_1 = \left\{\beta \in \mathcal{H}_{weak} : \beta^{(1)} = (\underbrace{1,...,1}_{r_1},\underbrace{0,...,0}_{p_n - r_1}), \left\|\beta^{(2)}\right\|_0 = r_2\right\}$$

and

$$\mathcal{H}_2 = \left\{\beta \in \mathcal{H}_{weak} : \left\|\beta^{(1)}\right\|_0 = r_1, \left\|\beta^{(2)}\right\|_0 = 0\right\}.$$

Similar metric entropy results of the above two subsets can be derived in the same fashion as in Lemmas 1 and 2. Other arguments are the same as in the proof of ((4.5)). □

C.3 Proof of Theorem 9

Proof C.4

For the upper bound under no heredity, we consider the model $\hat{\mathtt{I}} = \arg\min_{\mathtt{I}\in\mathtt{I}^{no}_{r_1,r_2}} \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{\mathtt{I}})^2$ with the model descriptive complexity $C_{\mathtt{I}^{no}_{r_1,r_2}} =$

$\log\binom{p_n}{r_1} + \log\binom{\binom{p_n}{2}}{r_2}$ for $1 \le r_1 \le p_n \wedge n$ and $0 \le r_2 \le \binom{p_n}{2} \wedge n$. The ABC criteria for the models are defined as in ((4.10)).

For the lower bound under no heredity, we consider the set

$$\mathcal{H}_{no} = \left\{\beta \in \ddot{\mathbb{R}}^{p_n} : \beta \in \{-1, 0, 1\}^{p_n + \binom{p_n}{2}}, \left\|\beta^{(1)}\right\|_0 \le r_1, \left\|\beta^{(2)}\right\|_0 \le r_2\right\}.$$

Then the two important subsets are instead

$$\mathcal{H}_1 = \left\{\beta \in \mathcal{H}_{no} : \beta^{(1)} = (\underbrace{1, ..., 1}_{r_1}, \underbrace{0, ..., 0}_{p_n - r_1}), \left\|\beta^{(2)}\right\|_0 = r_2\right\}$$

and

$$\mathcal{H}_2 = \left\{\beta \in \mathcal{H}_{no} : \left\|\beta^{(1)}\right\|_0 = r_1, \left\|\beta^{(2)}\right\|_0 = 0\right\}.$$

Similar metric entropy results of the above two subsets can be derived in the same fashion as Lemmas 1 and 2.

Other arguments are the same as in the proofs of ((4.4)) and ((4.5)). □

C.4 Proof of Theorem 10

The model descriptive complexity term $\lambda\sigma^2 C_{\mathrm{I}}$ plays a fundamental role in model selection theory Barron and Cover (1991); Barron et al. (1999); Yang (1999); Wang et al. (2014). Since we are considering models with interaction terms, the model descriptive complexity C_{I} reflects our comprehension of the model complexity other than the total number of parameters only. The detailed designation of the descriptive complexity usually depends on the class of models of interest. Instead of interpreting C_{I} as the code length (or description length) of describing the model index, one can also treat $\exp(-C_{\mathrm{I}})$ as the prior probability assigned to the model from a Bayesian viewpoint.

Proof C.5

The candidate set can be represented as the union of the candidate sets under three heredity conditions, i.e., $\bar{\mathcal{F}} = \mathcal{F}_{strong} \cup \mathcal{F}_{weak} \cup \mathcal{F}_{no}$, with

$$\mathcal{F}_{strong} := \{\mathtt{I}_{p_n,(p_n^2-p_n)/2}\} \cup \{\mathtt{I}^{strong}_{k_1,k_2}\},$$

$$\mathcal{F}_{weak} := \{\mathtt{I}_{p_n,(p_n^2-p_n)/2}\} \cup \{\mathtt{I}^{weak}_{k_1,k_2}\},$$

$$\mathcal{F}_{no} := \{\mathtt{I}_{p_n,(p_n^2-p_n)/2}\} \cup \{\mathtt{I}^{no}_{k_1,k_2}\}.$$

When $h \in \mathcal{F}_0(r_1, r_2; \ddot{\mathbb{R}}^{p_n}_{strong})$, there exists a specific model in $\mathcal{F}_{strong}$ such that the projection estimator of this model is equal to $\mathbf{h}$. Also, the projection of $\mathbf{h}$ onto the full design matrix is still $\mathbf{h}$. We denote the two models as $\mathtt{I}_{r_1,r_2}$ and $\mathtt{I}_{p_n,(p_n)(p_n-1)/2}$ respectively. It follows that

$$\begin{aligned}
E(\mathcal{L}(\hat{Y}^{\bar{\mathcal{F}}})) &\leq c \inf_{\mathtt{I}\in\bar{\mathcal{F}}} \left(\frac{1}{n} \left\| \mathbf{h}_{\mathtt{I}} - \mathbf{h} \right\|_2^2 + \frac{\sigma^2 r_{\mathtt{I}}}{n} + \frac{\lambda\sigma^2 C_{\mathtt{I}}}{n} \right) \\
&\leq c \inf_{\mathtt{I}\in\mathcal{F}_{strong}} \left(\frac{1}{n} \left\| \mathbf{h}_{\mathtt{I}} - \mathbf{h} \right\|_2^2 + \frac{\sigma^2 r_{\mathtt{I}}}{n} + \frac{\lambda\sigma^2 C_{\mathtt{I}}}{n} \right) \qquad \text{(C.2)} \\
&\leq c \left(\left\| \mathbf{h}_{\mathtt{I}_{r_1,r_2}} - \mathbf{h} \right\|^2 + \frac{\sigma^2 r_{\mathtt{I}_{r_1,r_2}}}{n} + \frac{\lambda\sigma^2 C_{\mathtt{I}_{r_1,r_2}}}{n} \right) \\
&\quad \wedge c \left(\left\| \mathbf{h}_{\mathtt{I}_{p_n,p_n(p_n-1)/2}} - \mathbf{h} \right\|^2 + \frac{\sigma^2 R_{\mathbf{Z}}}{n} + \frac{-\lambda\sigma^2 \log\pi_0}{n} \right) \\
&= \underbrace{\frac{c}{n}\left(\sigma^2 r_{\mathtt{I}_{r_1,r_2}} + \lambda\sigma^2 C_{\mathtt{I}_{r_1,r_2}}\right)}_{(i)} \wedge \underbrace{\frac{c}{n}\left(\sigma^2 R_{\mathbf{Z}} - \lambda\sigma^2 \log\pi_0\right)}_{(ii)}, \qquad \text{(C.3)}
\end{aligned}$$

where $R_{\mathbf{Z}}$ is the rank of the full design matrix, the first inequality follows from ((C.1)), the second inequality follows from $\mathcal{F}_{strong} \subseteq \bar{\mathcal{F}}$ and the third inequality results from the evaluation of ((C.2)) at $\mathtt{I}_{r_1,r_2}$ and $\mathtt{I}_{p_n,p_n(p_n-1)/2}$. The two terms (i) and (ii) are

bounded as follows:

$$\begin{aligned}(i) &\leq \frac{c_1\lambda}{n}\sigma^2\left(\frac{r_1+r_2}{\lambda} - \log\pi_1 + \log p_n + \log\binom{r_1}{2} + \log\binom{p_n}{r_1} + \log\binom{\binom{r_1}{2}}{r_2}\right)\\ &\leq \frac{c_1\lambda}{n}\sigma^2\left(\frac{r_1+r_2}{\lambda} - \log\pi_1 + r_1(1+\log\frac{p_n}{r_1})\right.\\ &\quad \left. + \log r_1^2 + r_1(1+\log\frac{p_n}{r_1}) + r_2(1+\log\frac{\binom{r_1}{2}}{r_2})\right)\\ &\leq \frac{c_2}{n}\sigma^2\left(r_1(1+\log\frac{p_n}{r_1}) + r_2(1+\log\frac{\binom{r_1}{2}}{r_2})\right),\end{aligned}$$

and

$$\begin{aligned}(ii) &\leq \frac{c}{n}\left(\sigma^2 R_{\mathbf{Z}} - \lambda\sigma^2\log\pi_0\right)\\ &\leq \frac{c_3}{n}\sigma^2 R_{\mathbf{Z}}.\end{aligned}$$

Therefore, we have

$$E(\mathcal{L}(\hat{Y}^{\bar{\mathcal{F}}})) \leq \frac{\max(c_2,c_3)\cdot\sigma^2}{n}\left[\left(r_1(1+\log\frac{p_n}{r_1}) + r_2(1+\log\frac{\binom{r_1}{2}}{r_2})\right) \wedge R_{\mathbf{Z}}\right],$$

where c_1, c_2, c_3 are some constants that depend only on the constant λ. Thus the desired minimax upper bound follows.

When $h \in \mathcal{F}_0(r_1,r_2;\ddot{\mathbb{R}}^{p_n}_{weak})$ or $h \in \mathcal{F}_0(r_1,r_2;\ddot{\mathbb{R}}^{p_n})$, with $\mathtt{I} \in \mathcal{F}_{weak}$ or $\mathtt{I} \in \mathcal{F}_{no}$ replacing $\mathtt{I} \in \mathcal{F}_{strong}$ in ((C.2)), the quantity (i) in ((C.3)) will instead be no greater than

$$\frac{c_1\lambda}{n}\sigma^2\left(\frac{r_1+r_2}{\lambda} - \log\pi_2 + \log p_n + \log K + \log\binom{p_n}{r_1} + \log\binom{K}{r_2}\right)$$

with $K = r_1 p_n - \binom{r_1}{2} - r_1$ under weak heredity $h \in \mathcal{F}_0(r_1,r_2;\ddot{\mathbb{R}}^{p_n}_{weak})$, or

$$(i) \leq \frac{c_1\lambda}{n}\sigma^2\left(\frac{r_1+r_2}{\lambda} - \log\pi_3 + \log p_n + \log\binom{p_n}{2} + \log\binom{p_n}{r_1} + \log\binom{\binom{p_n}{2}}{r_2}\right)$$

under no heredity $h \in \mathcal{F}_0(r_1,r_2;\ddot{\mathbb{R}}^{p_n})$. The different constants π_2, π_3 does not affect

the conclusion in terms of order. Following the same arguments in the proof of strong heredity, the desired results follow when the underlying heredity condition is weak heredity or no heredity. □

C.5 An example when SRC is not satisfied

For simplicity, let us consider an example where the regression mean function includes only one main effect term, i.e., $r_1 = 1, r_2 = 0$. The corresponding SRC assumption with $l_1 = r_1 = 1, l_2 = r_2 = 0$ will be that there exist constants $b_1, b_2 > 0$ (not depend on n) such that for any $\beta \in \mathbb{R}^{p_n}$ with $\|\beta\|_0 \leq 2$, we have

$$b_1 \|\beta\|_2 \leq \frac{1}{\sqrt{n}} \|\mathbf{Z}\beta\|_2 \leq b_2 \|\beta\|_2 , \tag{C.4}$$

where the design matrix $\mathbf{Z} = \mathbf{X}$ is the matrix that contains the main effects.

Assume the first $R_{\mathbf{Z}}$ columns of $\mathbf{Z}$ are linearly independent and denote $\mathbf{Z} = (\mathbf{Z}^1, \mathbf{Z}^2)$, where $\mathbf{Z}^1 = (\mathbf{Z}_1, ..., \mathbf{Z}_{R_{\mathbf{Z}}})$ is the $n \times R_{\mathbf{Z}}$ submatrix with $rank(\mathbf{Z}^1) = R_{\mathbf{Z}}$. Suppose the submatrix $\mathbf{Z}^1$ satisfies the SRC assumption. Assume that $\|\mathbf{Z}_i\|_2 = f(n)$ for $1 \leq i \leq p_n$. For the purpose of illustration, we set $f(n) = \sqrt{n}$.

Let A be the collection of all columns in $\mathbf{Z}^2$, then A is a subset of $\{z | z = \mathbf{Z}^1\alpha, \alpha \in \mathbb{R}^{R_{\mathbf{Z}}}, \|z\|_2 = f(n)\}$. Then A should satisfy that $\forall z, z' \in A$, we have $b_1 \leq \frac{1}{\sqrt{n}} \|a_1 z + a_2 z'\|_2 \leq b_2$ for all $a_1, a_2 \in \mathbb{R}$ and $a_1^2 + a_2^2 = 1$. We know

$$\frac{1}{\sqrt{n}} \|a_1 z + a_2 z'\|_2 = \frac{1}{\sqrt{n}} \sqrt{a_1^2 \|z\|_2^2 + a_2^2 \|z'\|_2^2 + 2a_1 a_2 \|z\|_2 \|z'\|_2 \cos\theta},$$

where θ is the angle between the two n-dimensional vectors z and z'.

Thus we have

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sqrt{a_1^2\left\|z\right\|_2^2+a_2^2\left\|z'\right\|_2^2+2a_1a_2\left\|z\right\|_2\left\|z'\right\|_2\cos\theta} \\ = \; & \frac{f(n)}{\sqrt{n}}\sqrt{a_1^2+a_2^2+2a_1a_2\cos\theta} \\ = \; & \sqrt{1+2a_1a_2\cos\theta}. \end{aligned}$$

Then $\sqrt{1+2a_1a_2\cos\theta} \geq b_1$ for all $a_1^2+a_2^2=1$ (otherwise $\frac{1}{\sqrt{n}}\left\|a_1z+a_2\mathbf{Z}_i\right\|_2$ is less than b_1, which violates the SRC assumption). Since $-1 \leq 2a_1a_2 \leq 1$ for $a_1^2+a_2^2=1$, we have $b_1 \leq \sqrt{1-|\cos\theta|}$, which implies $|\cos\theta| \leq 1-b_1^2$. Thus, we have

$$b_1 \leq \frac{1}{\sqrt{n}}\left\|a_1z+a_2z'\right\|_2 = \sqrt{1+2a_1a_2\cos\theta} \leq \sqrt{1+|\cos\theta|} \leq \sqrt{1+1-b_1^2}. \quad \text{(C.5)}$$

By setting $a_1=a_2=\frac{1}{\sqrt{2}}$ in ((C.5)), the the pairwise l_2 distance between any two elements z, z' in A should satisfy $\sqrt{2}b_1 \leq \frac{1}{\sqrt{n}}\|z-z'\|_2 \leq \sqrt{4-2b_1^2}$. It is well known that the ϵ-covering entropy of the $R_{\mathbf{Z}}$-dimensional unit ball $\mathbb{B}$ is of order $R_{\mathbf{Z}}\log(1/\epsilon)$. We denote $\sqrt{n}\mathbb{B}$ as a ball of radius $\sqrt{n}$. Let $\epsilon=\sqrt{2n}b_1/2$, there exists a positive constant c_1 such that $\log N(\epsilon;\sqrt{n}\mathbb{B},l_2) \leq c_1R_{\mathbf{Z}}\log(\sqrt{n}/\epsilon) = c_1R_{\mathbf{Z}}\log(\sqrt{2}/b_1)$. Since A is a 2ϵ-packing set of a ball of radius $f(n)=\sqrt{n}$, its cardinality satisfies $\log|A| \leq \log M(2\epsilon;\sqrt{n}\mathbb{B},l_2)$. The covering number and the packing number are closely related as in the well-known inequality $M(\epsilon;\mathcal{X},\rho) \leq N(\frac{\epsilon}{2};\mathcal{X},\rho) \leq M(\frac{\epsilon}{2};\mathcal{X},\rho)$. Thus we have $\log|A| \leq \log M(2\epsilon;\sqrt{n}\mathbb{B},l_2) \leq \log N(\epsilon;\sqrt{n}\mathbb{B},l_2) \leq c_1R_{\mathbf{Z}}\log(\sqrt{2}/b_1)$, which implies A has at most $(\sqrt{2}/b_1)^{c_1R_{\mathbf{Z}}}$ elements under the SRC assumption. Thus, as long as $p_n > (\sqrt{2}/b_1)^{c_1R_{\mathbf{Z}}}$, the SRC assumption will not be satisfied because the SRC assumption requires that ((C.4)) must hold for any pair of columns in $\mathbf{Z}$. In this case, the lower bound $r_1(1+\log(p_n/r_1))/n$ in our theorems does not apply.