Published in final edited form as:
J Am Stat Assoc. 2014 ; 109(507): 1285–1301. doi:10.1080/01621459.2014.881741.

Interaction Screening for Ultra-High Dimensional Data

Ning Hao and
Assistant Professor, Department of Mathematics, University of Arizona, Tucson, AZ 85721

Hao Helen Zhang
Associate Professor, Department of Mathematics, University of Arizona, Tucson, AZ 85721

Ning Hao: nhao@math.arizona.edu; Hao Helen Zhang: hzhang@math.arizona.edu

Abstract

In ultra-high dimensional data analysis, it is extremely challenging to identify important interaction effects, and a top concern in practice is computational feasibility. For a data set with n observations and p predictors, the augmented design matrix including all linear and order-2 terms is of size $n \times (p^2 + 3p)/2$. When p is large, say more than tens of hundreds, the number of interactions is enormous and beyond the capacity of standard machines and software tools for storage and analysis. In theory, the interaction selection consistency is hard to achieve in high dimensional settings. Interaction effects have heavier tails and more complex covariance structures than main effects in a random design, making theoretical analysis difficult. In this article, we propose to tackle these issues by forward-selection based procedures called iFOR, which identify interaction effects in a greedy forward fashion while maintaining the natural hierarchical model structure. Two algorithms, iFORT and iFORM, are studied. Computationally, the iFOR procedures are designed to be simple and fast to implement. No complex optimization tools are needed, since only OLS-type calculations are involved; the iFOR algorithms avoid storing and manipulating the whole augmented matrix, so the memory and CPU requirement is minimal; the computational complexity is *linear* in p for sparse models, hence feasible for $p \gg n$. Theoretically, we prove that they possess sure screening property for ultra-high dimensional settings. Numerical examples are used to demonstrate their finite sample performance.

Keywords

Forward selection; Sure Screening; Heredity condition; Interaction; GWAS

1 Introduction

Ultra-high dimensionality is a significant feature of data collected in contemporary scientific research, owing to rapid advances of technologies and computer power. Big data are abundant in many areas including biology, genetics, medicine, finance, social science, environmental science, and so on. One major challenge in dealing with big data sets is that, the number of predictors p is much larger than the sample size n. In this paper, we allow p to be as large as $O(\exp(n^{\xi}))$ for some $\xi \in (0,1/2)$, which is described as *nonpolynomial* (NP)

Correspondence to: Hao Helen Zhang, hzhang@math.arizona.edu.

dimensionality in Fan & Song (2010). To extract useful information from such data and build an interpretable model with high prediction power, variable selection or screening must be employed. A variety of variable selection methods have been developed and in common use, such as the LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), Dantzig selector (Candes & Tao, 2007), elastic net (Zou & Hastie, 2005), minimax concave penalty (MCP) (Zhang, 2010), and others (Zou, 2006; Zou & Li, 2008). Many methods possess favorable theoretical properties such as model selection consistency (Zhao & Yu, 2006) and oracle properties (Fan & Lv, 2011). When p is much larger than n, sure screening is a more realistic goal to achieve than oracle properties or selection consistency (Fan & Lv, 2008; Wang, 2009). Sure screening assures that all important variables are identified with a probability tending to one, hence achieving effective dimension reduction without information loss and providing a reasonable starting point for low-dimensional methods to be applied.

Most existing methods for variable selection are designed for selecting main effects only. However, main effects may not be sufficient to characterize the relationship between the response and predictors in complex situations, where predictors work together. Interaction models provide a better approximation to the response surface, improve prediction accuracy, and bring new insight on the interplay between predictors. They are useful in social, political, and economic problems to identify non-trivial interactions between covariates in modeling election results, product sales, social networks, stock market changes. One interesting application is to study the effects of combinations of various behaviors and exposures on disease rates, commonly needed in bioassay and epidemiology. In genome-wide association studies (GWAS), there is growing interest to identify the interaction (epistatic) effects of SNPs (Evans et al., 2006; Manolio & Collins, 2007; Kooperberg & LeBlanc, 2008; Cordell, 2009), since gene-gene interactions may provide critical insight on the complex biological pathways that underpin human diseases. A common class of linear models considering two-way interactions assume

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\beta_{11}X_1^2+\beta_{12}X_1X_2+\cdots+\beta_{pp}X_p^2+\varepsilon, \quad (1.1)$$

where Y is the response, $X_1, \cdots, X_p$ are covariates, and ε is the error. Marginality principle (Nelder, 1977, 1994; McCullagh & Nelder, 1989; McCullagh, 2002) or heredity conditions (Hamada & Wu, 1992; Chipman, 1996; Chipman et al., 1997) are generally employed to characterize the hierarchical structure between main and interaction effects. In particular, the strong heredity condition is

$$\beta_{k\ell} \neq 0 \Rightarrow \beta_k\beta_\ell \neq 0,$$

i.e., X_kX_ℓ is important only if its both parents X_k and X_ℓ are important. The weak heredity is

$$\beta_{k\ell} \neq 0 \Rightarrow \beta_k^2+\beta_\ell^2 \neq 0,$$

i.e., $X_k X_\ell$ is important only if at least one of X_k and X_ℓ is important.

Interaction selection for (1.1) has lately drawn much attention in the literature. Recent works include Efron et al. (2004), Turlach (2004), Yuan et al. (2007), Yuan et al. (2009), Zhao et al. (2009), and Choi et al. (2010), among others. In particular, Efron et al. (2004), Turlach (2004), and Yuan et al. (2007) considered enforcing the strong heredity principle in the LARS; Yuan et al. (2009) incorporated the structural relationship by imposing linear inequality constraints on coefficients; Zhao et al. (2009) introduced the Composite Absolute Penalties (CAP) to achieve hierarchy in variable selection. Choi et al. (2010) employed a special reparametrization of regression coefficients to enforce the heredity constraint. These procedures, except Efron et al. (2004), can be described as *joint analysis*, as they consider main and interaction effects in (1.1) altogether and make a global search over all candidate models. When p is small or moderate, joint analysis is effective in identifying important interaction effects. Some joint-analysis methods can produce consistency selection results under the strong heredity condition for a fixed p (**?**). However, joint-analysis methods become infeasible if p is very large. Two major limiting factors are memory requirement and computational cost. Joint analysis typically requires to store the entire augmented design matrix of size $n \times (p^2 + 3p)/2$. Take an example of $n = 200$, $p = 10,000$, where the total number of entries is $\approx 10^{10}$ and beyond the capacity of standard software such as R and MATLAB. Since sophisticated programming tools are needed to handle complex penalty structures (Zhao et al., 2009; Choi et al., 2010) or multiple inequality constraints (Yuan et al., 2009), joint analysis implementation can be extremely expensive. Furthermore, it is not clear whether selection consistency would still hold in ultra-high dimensional settings.

An alternative interaction selection tool is *two-stage* analysis: first select main effects only (by intentionally leaving interaction terms out) at Stage 1, then select interactions of main effects identified at Stage 1. When the data dimension is very large, two-stage approaches are possibly only feasible choices for practitioners (Wu et al., 2009, 2010). Despite their computational advantages over joint analysis, two-stage procedures have been criticized for their validity, even for low-dimensional data with $p < n$ (Turlach, 2004).

Motivated by the above practical and theoretical concerns, we propose new greedy-type model selection procedures for high dimensional interaction selection, study their numerical properties and performance, and provide rigorous theoretical justifications. In particular, we consider *i*nteraction selection procedures featured with *FOR*ward selection, which are referred to *iFOR*. Forward selection (FS) is a classical variable selection method in linear regression and it builds the model sequentially by adding one variable at a time. FS is easy to implement as it involves only simple OLS-type operations. Though the local search is sub-optimal, it is a necessary compromise when dealing with high dimensionality for the sake of computation. In this article, we propose two algorithms: iFORT and iFORM. The iFORT is a two-stage procedure: at the first stage, it selects only main effects (all quadratic terms and interactions ignored) by FS; at the second stage, interaction terms generated under the heredity condition are considered. The iFORM, on the other hand, selects main effects and interactions altogether in an iterative fashion. Compared to joint analysis procedures, the iFOR methods can incorporate the strong or weak heredity condition in a much simpler fashion. Their implementation does not require the storage of the entire augmented matrix, making them feasible for large problems. The memory and computational complexity are shown to be *linear* in p. In one simulation example with p = 10, 000 and n = 400, it takes iFOR fewer than 30 seconds to complete the selection process. Numerical examples suggest promising performance of iFOR in terms of effective coverage. In additional to the new algorithms and numerical results, another major goal of this work is to investigate theoretical properties of iFORT and understand their asymptotic behaviors. By rigorously analyzing the covariance structure between main effects and interaction terms, we prove that the iFORT has a sure screening property for ultra-high dimensional settings. This is the first theoretical justification of two-stage approaches.

The rest of this article is organized as follows. Section 2 introduces the basic model setup and the new procedures: iFORT and iFORM, under the strong heredity condition. Major theoretical results are presented in Section 3. Section 4 extends the iFOR to the context of the weak heredity condition. Numerical results are demonstrated in Sections 5 and .6 Final remarks are given in Section 7. All technical proofs are relegated to the Appendix.

2 Methodology

2.1 Model Setup and Notations

Given n IID observations $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$, we consider a regression model with linear and second-order terms

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}^{(1)} + \mathbf{z}_i^\top \boldsymbol{\beta}^{(2)} + \varepsilon_i, \quad 1 \le i \le n, \quad (2.1)$$

where Y_i is a real-valued response, $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})$ is a p-dimensional vector, the vector $\mathbf{z}_i = (X_{i1}^2, X_{i1}X_{i2}, \ldots, X_{i1}X_{ip}, X_{i2}^2, X_{i2}X_{i3}, \ldots, X_{ip}^2)^\top$ contains quadratic and two-way interaction terms, β_0 is the intercept, $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are respectively regression coefficients of linear effects and order-2 effects, and ε_i is the noise with mean zero and finite variance σ^2. The length of $\mathbf{z}_i$ or $\boldsymbol{\beta}^{(2)}$ is $q = (p + p^2)/2$. The entire parameter vector is $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)\top}, \boldsymbol{\beta}^{(2)\top})^\top$. Throughout this article, we assume that $E(X_{ij}) = 0$, $\mathrm{Var}(X_{ij}) = 1$, $E(Y_i) = 0$, $\mathrm{Var}(Y_i) = 1$ in (2.1) for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We also assume that all the quadratic effects and two-way interactions are centered, i.e., $\mathbf{z}_i = (\ldots, X_{ik}X_{i\ell} - E(X_{ik}X_{i\ell}), \ldots)^\top$. This eliminates the need of the intercept term β_0 in (2.1).

For convenience, denote $([\mathbf{x}_i^\top])_{i=1}^n$ as the design matrix containing only linear effects, and $([\mathbf{x}_i^\top, \mathbf{z}_i^\top])_{i=1}^n$ as the augmented design matrix. Define the index sets of linear and order-2 terms as

$$\mathscr{P}_1 = \{1, 2, \ldots, p\}, \quad \mathscr{P}_2 = \{(k, \ell): 1 \le k \le \ell \le p\}.$$

In (2.1), any term $\beta_j \neq 0$ or $\beta_{jk} \neq 0$ is regarded as *relevant*; the corresponding predictor can be a linear, or quadratic, or interaction effect. We define the nonzero linear and order-2 effects as

$$\mathscr{T}_1\{j:\beta_j \neq 0, j \in \mathscr{P}_1\}, \quad \mathscr{T}_2=\{(j,k):\beta_{jk} \neq 0, (j,k) \in \mathscr{P}_2\}.$$

The *full* model is $\mathcal{F} = \mathcal{P}_1 \cup \mathcal{P}_2$ and the *true* model $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$ For any model $\mathcal{M}$, use $|\mathcal{M}|$ to denote its model size, i.e., the number of predictors contained in $\mathcal{M}$. We have $|\mathcal{P}_1| = p$, $|\mathcal{P}_2| = q$, and $|\mathcal{F}| = d = p+q$. We assume $|\mathcal{T}_1| = p_0$ and $|\mathcal{T}_2| = q_0$, and then the true model size$|\mathcal{T}| = d_0 = p_0 + q_0$. In the literature, variable selection for (2.1) have been studied by penalized least squares using the augmented matrix $([\mathbf{x}_i^\top, \mathbf{z}_i^\top])_{i=1}^n$ as the covariates and conducting variable selection under heredity principles. They work quite well when p is moderate. But when p is big, their implementation becomes infeasible since the full model size$d$ increases quadratically in p. For example, p = 50 and d = 1, 325, p = 500 and d = 125, 750, and p = 5, 000 and d = 12, 507, 500.

Next we give a review of the FS solution path algorithm (Wang, 2009), which is closely related to the interaction selection algorithms under consideration. For each $1 \le k \le n$, we use $\mathcal{S}_k$ to denote the index of selected variables at the end of the kth step. Let $\mathrm{RSS}_{\mathcal{M}}$ be the residual sum of squares (RSS) using model $\mathcal{M}$ to fit the data.

Forward Selection (FS)

Initial step: Set $k = 0$ and $\mathcal{S}_0 = \emptyset$.

Iterative step $k = k + 1$: If $k > n$, stop. Otherwise, given $\mathcal{S}_{k-1}$, for every $j \in \mathcal{P}_1 \setminus \mathcal{S}_{k-1}$, construct a candidate model $\mathcal{M}_{j,k-1} = \mathcal{S}_{k-1} \cup \{j\}$. Compute the $\mathrm{RSS}_{\mathcal{M}_{j,k-1}}$ for each j. Find a_k = arg $\min_{j \in \mathcal{P}_1 \setminus \mathcal{S}_{k-1}} \mathrm{RSS}_{\mathcal{M}_{j,k-1}}$ and update $\mathcal{S}_k = \mathcal{S}_{k-1} \cup \{a_k\}$. Repeat this step until stop.

The FS algorithm produces a solution path consisting of n nested models $\mathcal{S}_1 \subset, \cdots \subset \mathcal{S}_n$, where $\mathcal{S}_k = \{a_1, \cdots, a_k\}$ for $1 \le k \le n$. When $p \gg n$, the FS automatically terminates after n steps when RSS reduces to zero. Since the solution path of the FS depends only on the subspaces spanned by the predictor vectors (column vectors in the design matrix), centering and standardization does not change the solution path. Wang (2009) showed the screening consistency of the FS for main-effect selection under the ultra-high dimensional setup.

One straightforward way of extending the FS to the interaction selection is to apply FS directly to model (2.1), ignoring the hierarchical structure. We name this procedure FS2 to distinguish it from the usual FS for main effect selection. Based on our empirical experience, FS2 works well for small and moderate p in sparse settings. In Section 3, we prove that FS2 has a sure screening property for interaction selection under some regularity conditions. However, similar to joint-analysis methods, the implementation of FS2 requires to store the entire augmented design matrix or call the features repeatedly during computation procedure, making it difficult for high dimensional data analysis.

2.2 New Methods: iFOR

We propose two forward selection based algorithms for interaction selection. The new algorithms naturally incorporate the marginality or heredity principles (Zhao et al., 2009; Yuan et al., 2009; Choi et al., 2010), without invoking complex constraint or optimization tools as done in joint analysis. Throughout this section, we use $\mathcal{C}$ to denote the *candidate* index set which consists of all the terms to be considered for selection in the immediately following step.

We first describe the *t*wo-stage approach (iFORT) algorithm. At Stage 1, only main effects are selected by FS, all of the order-2 terms left out of the model. Denote the selected main-effect set by $\hat{\mathcal{M}}$ At Stage 2, we expand $\hat{\mathcal{M}}$ by adding all the two-way interactions within $\hat{\mathcal{M}}$ and then implement FS on the expanded set while forcing $\hat{\mathcal{M}}$ to stay in the final model.

Two-stage iFOR (iFORT)

Stage 1: Define $\mathcal{C} = \mathcal{P}_1$. Implement FS on $\mathcal{C}$. The resulting solution path is $\{\mathscr{S}_t^{(1)}, t = 1, 2, \cdots\}$, and the selected main effects are $\hat{\mathcal{M}} = \{j_1, \ldots, j_{t_1}\}$.

Stage 2: Update $\mathcal{C} = \hat{\mathcal{M}} \cup \{(k, l) : k \in \hat{\mathcal{M}} \text{ and } l \in \hat{\mathcal{M}}\}$. Implement FS on $\mathcal{C}$ by forcing-in $\hat{\mathcal{M}}$. Denote the solution path by $\{\mathscr{S}_{t_1+t}^{(2)}, t = 1, 2, \cdots\}$.

The iFORT is simple, fast, and feasible to implement for high dimensional data analysis. It does not require complex optimization tools, and the strong heredity condition is automatically satisfied in the final model by forcing-in $\hat{\mathcal{M}}$. If the model is sparse, the number of important linear effects p_0 would be small, so the number of terms considered at Stage 2 would be much smaller than $(p^2 + p)/2$. Theoretical properties of iFORT are studied in Section 3.

The iFORT separately selects main effects and order-2 terms at two stages. Alternatively, one may select them altogether under the marginality principle, and this leads to the algorithm iFORM. The main idea of the iFORM is to apply FS to a sub-model of model (2.1) indexed by a dynamic candidate set $\mathcal{C}$. At step *t*, we use $\mathcal{S}_t$, $\mathcal{M}_t$ and $\mathcal{C}_t$ respectively to represent the index set of all selected effects, selected main effects, and current candidate set. Initially, $\mathcal{C} = \mathcal{P}_1$, i.e., all the main effects. Then the candidate set $\mathcal{C}$ grows gradually by adding two-way interactions between the main effects already in the model. In other words, we update $\mathcal{C}_t$ by defining $\mathcal{C}_t = \mathcal{P}_1 \cup \{(k, \ell) : k, \ell \in \mathcal{M}_t\}$.

iFOR under Marginality Principle (iFORM)

Step 1: (Initialization) Set $\mathcal{S}_0 = \emptyset$, $\mathcal{M}_0 = \emptyset$ and $\mathcal{C}_0 = \mathcal{P}_1$.

Step 2: (Selection) In the *t*th step with given $\mathcal{S}_{t-1}$, $\mathcal{C}_{t-1}$ and $\mathcal{M}_{t-1}$, forward regression is used to select one more predictor from $\mathcal{C}_{t-1} \backslash \mathcal{S}_{t-1}$ into the model. We add the selected one into $\mathcal{S}_{t-1}$ to get $\mathcal{S}_t$. We also update $\mathcal{C}_t$ and $\mathcal{M}_t$ if the newly selected predictor is a main effect. Otherwise, $\mathcal{C}_t = \mathcal{C}_{t-1}$ and $\mathcal{M}_t = \mathcal{M}_{t-1}$.

Step 3: (Solution path) Iterating Step 2, we get a solution path $\{\mathcal{S}_t: t = 1, 2, \ldots D\}$.

In the above algorithm, *D* is chosen as a reasonable upper bound of d_0 (the total number of important effects), to terminate the procedure. A direct advantage of the iFORM is that it allows the interactions to enter the model early, making it easier to select weak relevant main effects. Moreover, when we decide the optimal model along the solution path, we only need to use model size selection criteria, say BIC, once, while for iFORT, we have to use BIC twice which may cause additional error in practice even if the solution path is correct.

Our empirical experience also suggest the iFORM has better finite sample performance. The screening consistency of iFORM is shown in Section 3.3.

To select the optimal model from the FS path, we consider the use of BIC. There are two types of BIC proposed in the literature, the standard BIC

$$\text{BIC}_1(\hat{\mathscr{M}})=\log\hat{\sigma}^2_{\hat{\mathscr{M}}}+n^{-1}|\hat{\mathscr{M}}|\log(n)$$

and the BIC specially designed for high dimensional data (Chen & Chen, 2008)

$$\text{BIC}_2(\hat{\mathscr{M}})=\log\hat{\sigma}^2_{\hat{\mathscr{M}}}+n^{-1}|\hat{\mathscr{M}}|(\log(n)+2\log d^*)$$

where d^* is the number of predictors in the full model. The only difference between two BICs is the extra term 2 log d^* in BIC_2. Chen & Chen (2008) derived BIC_2 by controlling the false discovery rate (FDR) and showed that it is selection consistent if $d^* = O(n^{\xi})$ for some $\xi > 0$. Wang (2009) showed its selection consistency for FS under ultra-high dimensional setup $d^* = O(\exp(n^{\xi}))$. Since we deal with the ultra-high dimensional data, we use BIC_2 for iFORM and the first stage of iFORT. At the second stage of iFORT, since the number of candidate predictors is already dramatically reduced after the first stage, BIC_1 is more appropriate. Section 5 demonstrates their effective performance, in terms of coverage, false discovery control, and prediction accuracy.

2.3 Computational Complexity and Practical Issues

We show that the computational complexity of iFOR procedures is *linear* in p, which explains their feasibility for $p \gg n$. The FS algorithm described in Section 2.1 is equivalent to the following procedure. At each step, the response is regressed on the most correlated covariate, and the residual is calculated and used as the new response in next step. After the most correlated covariate (say, X_1) is selected, all other covariates are regressed on X_1, and then the covariates are substituted by the corresponding normalized residuals, which are used as the new covariates in next step. Note that the computation complexity of each step is $O(nm)$, where n is the sample size and m is the number of predictors in the candidate set. First, the absolute correlations between the response and all covariates in the current candidate set are calculated at each step, so the complexity is $O(nm)$. Once the most correlated covariate is selected, the response and all other covariates are regressed on it, whose cost is also $O(nm)$. For the iFORT and iFORM algorithms, the number of steps to build the whole solution path is at most n, so the number of main effects selected is not larger than n. This implies that, at each step, there are at most $p + n(n + 1)/2$ predictors in the candidate set, i.e., $m \le p + n(n + 1)$ holds for any step. Therefore, the overall complexity is $nO(n(p + n(n + 1)) = O(n^2 p + n^4)$, which is linear in p.

The parameter D controls the length of the solution path for the iFORM. Since the final model is chosen based on BIC by comparing all the models along the path, the final model select results is not sensitive to the exact value of D as long as it is reasonably large. In practice, though d_0 is unknown, it is reasonable to assume that d_0 is much smaller than n in high dimensional sparse regression problems (Fan & Lv, 2008). In our numerical study, we have tried $D = n/2$, $n/3$, $n/4$ and obtained the same results since $D > d_0$. In general, we suggest $D = n/2$.

3 Theoretical Results

We study theoretical properties of iFOR. In literature, a long-term concern about two-stage methods is their theoretical validity, as the main effect selection at Stage 1 is conducted under a misspecified working model. In Section 3.1, we first prove that the iFORT is able to capture all important main effects under ultra-high dimensional settings. This fundamental result provides rigorous justifications for two-stage methods. In Sections 3.2, we prove that iFORT can identify all important interactions consistently with probability tending to one under heredity conditions. The screening consistency of iFORM is shown in Section 3.3.

3.1 Screening Consistency of iFORT for Main Effects

Recall the *true model* $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$, where $\mathcal{T}_1 \subset \mathcal{P}_1$ and $\mathcal{T}_2 \subset \mathcal{P}_2$. For any square matrix A, denote its smallest and largest eigenvalues respectively by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ Denote the covariance matrices of main linear effects and interactions (i.e. all degree 2 monomials) respectively by $\Sigma^{(1)}$ and $\Sigma^{(2)}$. The total covariant matrix is Σ. The following regularity conditions are needed.

(C1) Normality: $X_{i1}, \ldots, X_{ip}$ are jointly normal and marginally standard normal. $\varepsilon_i \sim N(0, \sigma^2)$ is independent of $X_{i1}, \ldots, X_{ip}$.

(C2) Covariance Matrix: We assume that there exist two constants $0 < \tau_{\min} < \tau_{\max} < \infty$, such that $2\tau_{\min} < \lambda_{\min}(\Sigma^{(1)}) \leq \lambda_{\max}(\Sigma^{(1)}) < \tau_{\max}/2$.

(C3) Signal strength: We assume that $\|\beta\| \leq C_\beta$ for some positive constant C_β and $\beta_{\min} \geq \nu_\beta n^{-\xi_{\min}}$, where $\beta_{\min} = \min_{k \in \mathcal{T}} |\beta_k|$.

(C4) Dimensionality and sparsity: There exist positive constants ξ, ξ_0 and ν such that $\log p \leq \nu n^{\xi}$, $d_0 \leq \nu n^{\xi_0}$ and $\xi + 6\xi_0 + 12\xi_{\min} < 1$, $\xi < \frac{1}{2}$.

Remark 3.1: Conditions (C1) to (C4) are standard in the literature of ultra-high dimensional inference (Fan & Lv, 2008; Zhang & Huang, 2008). The normality assumption (C1) is extensively used in the past literature to facilitate proof (Fan & Lv, 2008; Zhang & Huang, 2008; Wang, 2009). (C2) requires the design matrix of main effects to be well-behaved. (C1) and (C2) together assure the Sparse Riesz Condition (Zhang & Huang, 2008); see the proof in Appendix for more details. (C3) requires that the smallest signal should not decay too fast, otherwise they can not be consistently identified; see (Fan & Peng, 2004) for more discussions. (C4) allows the dimension p to diverge with n at an exponential rate, or the NP dimensionality (Fan & Lv, 2008). Intuitively, one would expect that stronger conditions are needed to develop theory for interaction selection due to their heavier tails. However, to our satisfaction, conditions (C1) to (C4) are comparable to those used in the main-effect selection literature (Fan & Lv, 2008; Wang, 2009). The only difference is $\xi < \frac{1}{2}$ in (C4) while $\xi < 1$ is used in Wang (2009), due to heavier tails of interaction terms. Note if X_{1j} are sub-

Gaussian with $\mathrm{E}(e^{aX_{1j}^2})<b$ for positive constants *a* and *b*, typically, we can only bound a product term by $\mathrm{E}(e^{2aX_{1j}X_{1k}}) < b^2$.

Theorem 1: (sure screening of main effects) Define $K=2\tau_{\max}\nu C_\beta^2\tau_{\min}^{-2}\nu_\beta^{-4}$. Under conditions (C1)–(C4), the first stage of iFORT is screening consistent for the main effects. For $t_1 \geq K\nu n^{2\xi_0+4\xi_{\min}}$,

$$\mathbf{P}(\mathscr{T}_1 \subset \mathscr{S}_{t_1}^{(1)}) \to 1 \quad as \quad n \to \infty. \quad (3.1)$$

Next we give insight on why screening consistency (3.1) still holds for selection under a mis-specified model. A key observation is Lemma 1 in Appendix, which says, under (C1),

$$\sum = \begin{pmatrix} \sum^{(1)} & 0 \\ 0 & \sum^{(2)} \end{pmatrix}.$$

The block structure of Σ guarantees that ignored important interactions terms have minimal affects to the procedure at Stage 1. Imaging if there are some nonzero terms on the right top corner of Σ, we have to put some strong and complicated conditions on Σ to guarantee screening consistency.

Remark 3.2: In general, as long as Σ has a block structure, Theorem 1 holds even without normality. Here (C1) is used as a convenient and sufficient condition to assure the covariance block structure. There are other weaker but sufficient conditions (C1)′ or (C1)″, which can replace (C1):

(C1)′ X_{ij} is sub-Gaussian marginally, and their joint distribution is symmetric with respect to **0**.

(C1)″ X_{ij} is sub-Gaussian marginally, and their joint distribution has varnished third moments.

3.2 Screening Consistency of iFORT for Interaction Effects

After Stage 1, the iFORT essentially reduces the main effect dimensionality from *p* to $t_1=o(n^{\frac{1}{3}})$, which is significant if $p \gg n$. Using (C4), it is straightforward to show $2\xi_0+4\xi_{\min}<\frac{1}{3}$. Next we study the asymptotic behaviors of iFORT for interaction selection under the strong heredity:

(H1). Strong heredity condition: $\beta_{k\ell} \neq 0 \Rightarrow \beta_k\beta_\ell \neq 0$.

Under (H1), the interaction selection of iFORT at Stage 2 does not need to deal with high dimensional predictors any more, since the number of selected main effects is $o(n^{\frac{1}{3}})$. Even if include all interactions within the selected model at Stage 1, the final model has cardinality $o(n^{\frac{2}{3}})$. Corollary 1 gives the fundamental result: the iFORT is screening consistent for interaction selection under the heredity condition for ultra-high dimensional settings.

Corollary 1: (sure screening of interactions) Conditional on (3.1) and (H1), for $t_2 \geq K\nu n^{2\xi_0+4\xi_{\min}}$,

$$\mathbf{P}(\mathscr{T} \subset \mathscr{S}^{(2)}_{t_1+t_2}) \to 1 \quad as \quad n \to \infty.$$

Remark 3.3: The strong heredity is necessary to ensure the consistency of two-stage procedures for screening interaction terms. Otherwise, if X_1X_2 is important but neither X_1 nor X_2, then the main effects are not guaranteed to be identified at Step 1, and consequently, their interaction X_1X_2 might not be considered at the second step. We also point out that the strong heredity condition is actually not that strong with a simple illustration. Consider the case $p = 2$, where the full model space (for simplicity, ignoring two quadratic terms) can be represented by the parameter set $(\beta_0, \beta_1, \beta_2, \beta_{12})^\top$ in $\mathbb{R}^4$. The strong heredity condition covers the entire $\mathbb{R}^4$ except a couple of subsets, such as $\{\beta_0 = 0, \beta_1^2+\beta_2^2+\beta_{12}^2>0\}$ and $\{\beta_1\beta_2 = 0, \beta_{12}^2>0\}$. The excluded subsets have zero mass in $\mathbb{R}^4$, so the strong heredity condition is met by most models. This implies that the iFORT methods work for a generic model.

3.3 Screening Consistency of FS2 and iFORM

Naively, we can use any one-stage variable selection tool to fit (1.1) directly (as long as computation is feasible), ignoring the hierarchical structure. Though the model consistency or screening consistency result (Zhao & Yu, 2006; Wang, 2009; Fan & Lv, 2011) could be generalized to the context of interaction selection, the extension of earlier proofs is not straightforward due to heavy tails of interaction effects. Actually, all the existing proof technique would require some regularity conditions on the eigenvalues of $\Sigma^{(2)}$. Next, we establish the screening consistency of FS2 under conditions that are related only to $\Sigma^{(1)}$.

- **C2a** Covariance Matrix: Assume that there exist two constants $0<\tau_{\min}<\frac{1}{4}<1<\tau_{\max}<\infty$, such that $\sqrt{\tau_{\min}}<\lambda_{\min}(\sum^{(1)}) \leq \lambda_{\max}(\sum^{(1)})< \sqrt{\tau_{\max}/4}$.
- **C4a** Dimensionality and sparsity: There exist positive constants ξ, ξ_0 and ν, such that $\log p \leq \nu n^{\xi}$, $d_0 \leq \nu n^{\xi_0}$ and $\xi+6\xi_0+12\xi_{\min}<\frac{1}{2}$.

There is no essential difference between (C2a) and (C2). (C2a) is used only for easy presentation. (C4a) is slightly stronger than (C4). A remark is that under (C1) and (C2a), the population and sample covariance matrices Σ and $\hat{\Sigma}$ can be well controlled because $\Sigma^{(2)}$ can be explicitly represented by $\Sigma^{(1)}$. See Lemma 3 in the appendix. On the other hand, the screening consistency result below strongly depends on the normality condition (C1) since there is no easy way to capture the structure of $\Sigma^{(2)}$ by $\Sigma^{(1)}$ without normality condition.

Theorem 2: Under conditions (C1), (C2a), (C3), and (C4a), FS2 is screening consistent. For $t \geq K\nu n^{2\xi_0+4\xi_{\min}}$,

$$\mathbf{P}(\mathscr{T} \subset \mathscr{S}_t^{FS2}) \to 1 \quad as \quad n \to \infty.$$

The screening consistency of iFORM is implied in the proof of Theorem 2, as iFORM is similar to FS2 but with a restrictive candidate set each step.

Corollary 2: Under conditions (C1), (C2a), (C3), (C4a), and (H1), iFORM is screening consistent. For $t \geq K\nu n^{2\xi_0+4\xi_{\min}}$,

$$\mathbf{P}(\mathscr{T} \subset \mathscr{S}_t) \to 1 \quad as \quad n \to \infty.$$

4 Extensions to Weak Heredity

In some real applications, the weak heredity provides a useful alternative for the underlying model structure. Under the weak heredity, for a two-way interaction effect to be active, at least one of the parent effects need to be effective. In this section, we generalize the iFOR algorithms described in Section 2 to satisfy the weak heredity condition. Similar to the strong heredity situation, both iFOR algorithms under the weak heredity are easy to implement.

(H2). Weak heredity condition: $\beta_{k\ell} \neq 0 \Rightarrow \beta_k^2+\beta_\ell^2 \neq 0$.

iFORT Under Weak Heredity (iFORT-w)

Stage 1: Define $\mathcal{C} = \mathcal{P}_1$. Implement FS on $\mathcal{C}$. The resulting solution path is $\{\mathscr{S}_t^{(1)}, t = 1, 2, \cdots\}$, and the selected main effects are $\widehat{\mathcal{M}} = \{j_1, \ldots, j_{t_1}\}$.

Stage 2: Update $\mathcal{C} = \widehat{\mathcal{M}} \cup \{(k, l) : k \in \widehat{\mathcal{M}} \text{ or } l \in \widehat{\mathcal{M}}\}$. Implement FS on $\mathcal{C}$ by forcing-in $\widehat{\mathcal{M}}$. Denote the solution path by $\{\mathscr{S}_{t_1+t}^{(2)}, t = 1, 2, \cdots\}$.

For the iFORM extension, after selecting any new linear term, we need to expand the candidate set by including all of its interactions with the other linear effects. Denote by $\mathcal{M}_t$ the index set of selected linear effects at Step t. Under the weak heredity condition, we update $\mathcal{C}_t$ as

$$\mathscr{C}_t=\mathscr{P}_1 \cup \{(k,\ell):k \,\text{or}\, \ell \in \mathscr{M}_t\}.$$

For each t, we use $\mathcal{S}_t$, $\mathcal{M}_t$ and $\mathcal{C}_t$ to represent the index set of selected model, selected main effects and candidates set at Step t, respectively.

iFORM Under Weak Heredity (iFORM-w)

Step 1: (Initialization) Set $\mathcal{S}_0 = \emptyset$, $\mathcal{M}_0 = \emptyset$ and $\mathcal{C}_0 = \mathcal{P}_1$.

Step 2: (Selection) In the tth step with given $\mathcal{S}_{t-1}$, $\mathcal{C}_{t-1}$ and $\mathcal{M}_{t-1}$, forward regression is used to select one more predictor from $\mathcal{C}_{t-1} \setminus \mathcal{S}_{t-1}$ into the model. We add the selected one into $\mathcal{S}_{t-1}$ to get $\mathcal{S}_t$. We also update $\mathcal{C}_t$ and $\mathcal{M}_t$ if the newly selected predictor is a main effect. Otherwise, $\mathcal{C}_t = \mathcal{C}_{t-1}$ and $\mathcal{M}_t = \mathcal{M}_{t-1}$.

Step 3: (Solution path) Iterating Step 2, we get a solution path $\{\mathcal{S}_t : t = 1, 2, \ldots D\}$.

Remarks 4.1: The weak heredity condition is slightly more flexible than the strong heredity condition, and generally chooses a larger model. In practice, the weak heredity is more useful to identify important interactions with one weak parent effect (Yuan, Joseph, and Zou 2007). With regard to the computation speed, since the candidate set size at each step is larger than in the strong heredity case, the iFORT-w and iFORM-w are slower than the iFORT and iFORM.

5 Numerical Studies

5.1 Experiments and Setup

We demonstrate performance of the iFOR methods in various $p \gg n$ scenarios, including the regression settings with independent predictors, predictors with autoregressive (AR) correlation structure, compound symmetry (CS) correlation, and more complex settings as considered in Fan and Song (2010). We consider forward-based joint analysis (FS2), and the proposed forward-based procedures iFORT, iFORM, iFORT-w, iFORM-w. In literature, there are other two-step procedures which are not based on forward selection such as Mendel (Wu et al. 2009) and Screen and Clean (Wu et al. 2010). For comparison, we also include two such procedures, iMART1 and iMART2. The iMART1 screens main effects based on marginal correlation at Step 1, i.e., those that exceed a threshold are retained as candidate predictors, and then conducts the LASSO penalized regression on the expanded dictionary consisting of all the candidate predictors and their pairwise interaction terms at Step 2. The iMART2 first screens main effects by marginal correlation, then screens the pairwise products of the main effect candidates by pairwise correlation, and then implements the LASSO to obtain the final model. The standard BIC is used to select the tuning parameter of LASSO. The oracle (ORACL) procedure is also presented as the gold standard, which is generally not available in practice.

Recall that the full model is $\mathcal{F} = \mathcal{P}_1 \cup \mathcal{P}_2$, $|\mathcal{P}_1| = p$, $|\mathcal{P}_2| = q$. The true model is $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$, $|\mathcal{T}_1| = p_0$, $|\mathcal{T}_2| = q_0$. We run $M = 100$ Monte Carlo simulations and report their average performance in selecting linear effects and interactions, estimating coefficients, and making predictions. For the mth replication, let $\hat{\beta}^{(m)}$ denote the fitted regression coefficients, $\hat{\mathcal{T}}_1^{(m)}$ and $\hat{\mathcal{T}}_2^{(m)}$ respectively denote the selected linear effects and interactions. To evaluate linear effect selection, we report the

- Coverage probability(Cov) $\sum_{m=1}^{M} I(\mathcal{T}_1 \subset \hat{\mathcal{T}}_1^{(m)})/M$,
- Percentage of correct zeros (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j^{(m)}=0, \beta_j=0)/[M(p-p_0)]$,
- Percentage of incorrect zeros (Inc0): $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j^{(m)}=0, \beta_j \neq 0)/[Mp_0]$.

- Exact selection probability(Ext) $\sum_{m=1}^{M} I(\mathscr{T}_1=\hat{\mathscr{T}}_1^{(m)})/M$,

For interaction selection, we report

- Coverage probability (iCov) $\sum_{m=1}^{M} I(\mathscr{T}_2 \subset \hat{\mathscr{T}}_2^{(m)})/M$,
- Percentage of correct zeros (iCor0) $\sum_{m=1}^{M}\sum_{(j,k)\in \mathscr{P}_2} I(\hat{\beta}_{jk}^{(m)}=0, \beta_{jk}=0)/[M(q-q_0)]$,
- Percentage of incorrect zeros (iInc0) $\sum_{m=1}^{n}\sum_{(j,k)\in \mathscr{P}_2} I(\hat{\beta}_{jk}^{(m)}=0, \beta_{jk}\neq 0)/[Mq_0]$.
- Exact selection probability (iExt) $\sum_{m=1}^{M} I(\mathscr{T}_2=\hat{\mathscr{T}}_2^{(m)})/M$,

The overall model selection is measured by the model size $\sum_{m=1}^{M} |\hat{\mathscr{T}}_1^{(m)} \cup \hat{\mathscr{T}}_2^{(m)}|/M$. For estimation, we report the mean squared error (MSE) $\sum_{m=1}^{M}\sum_{j=1}^{p}(\hat{\beta}_j^{(m)}-\beta_j)^2+\sum_{(j,k)\in \mathscr{P}_2}(\hat{\beta}_{jk}^{(m)}-\beta_{jk})^2)/M$. For the prediction error, we report the out-of-sample R^2 (Rsq):

$$100\% \times \left\{1-\frac{\sum_{i=1}^{n}[Y_i^*-x_i^*\hat{\boldsymbol{\beta}}^{(1)}-z_i^*\hat{\boldsymbol{\beta}}^{(2)}]^2}{\sum_{i=1}^{n}(Y_i^*-\overline{Y}^*)^2}\right\},$$

where the test data ($\mathbf{X}_i^*, Y_i^*$), $i = 1, \cdots, n$ are generated independently from the same distribution as the training set, and $\overline{Y}^*=\frac{1}{n}\sum_{i=1}^{n} Y_i^*$. A larger Rsq suggests a better prediction. The standard error of Rsq is reported as well. We also report the average computation time.

5.2 Simulation Results

In all the examples, we generate the response Y from model (2.1) with $\sigma = 2, 3, 4$.

Example 1: (Independent predictors) Let $(n, p, p_0, q_0) = (100, 500, 4, 4)$. $\mathbf{X}$'s are iid from $MVN(\mathbf{0}, I_p)$. The true $\boldsymbol{\beta}^{(1)} = (3, 0, 3, 0, 0, 3, 0, 0, 0, 3, \mathbf{0}_{490})$, so $\mathscr{T}_1 = \{1, 3, 6, 10\}$. The important interaction set $\mathscr{T}_2 = \{(1, 3), (1, 6), (3, 10), (6, 10)\}$ with coefficient 2.

Example 2: (Autoregressive correlation) Consider the same setup as Example 1, except that $\mathbf{X}$ follows MVN with mean $\mathbf{0}$ and $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$ for $1 \le j, k \le p$.

Example 3: (High dimensional: AR) Let $(n, p, p_0, q_0) = (400, 5000, 10, 10)$. We generate $\mathbf{X}$ from MVN with mean 0 and $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$. The true $\boldsymbol{\beta}^{(1)} = (3, 3, 3, 3, 3, 2, 2, 2, 2, 2, \mathbf{0}_{4990})$. The nonzero interaction set is $\mathscr{T}_2 = \{(1, 2), (1, 3), (2, 3), (2, 5), (3, 4), (6, 8), (6, 10), (7, 8), (7, 9), (9, 10)\}$, and their coefficients are (2, 2, 2, 2, 2, 1, 1, 1, 1, 1).

Example 4: (High dimensional: AR) We increase the dimension $p = 10000$ in Example 3.

Example 5: (High dimensional: FS2010) We use the same setup as in Example 4, except that **X** has a more complex covariance structure as considered in Fan and Song (2010). First, we generate X_j, $j = 1, \cdots, 50$ independently from the standard normal distribution. Then we define

$$X_k = \sum_{j=1}^{s} X_j (-1)^{j+1}/5 + \sqrt{25-s}/5\varepsilon_k, \quad k=p-50, \cdots, p.$$

with $s = 10$ and $\{\varepsilon_k\}_{k=p-49}^{50}$ follow the standard normal distribution.

Example 6: (Weak Heredity) We use the same setup as in Example 3, except the nonzero interaction set $\mathcal{T}_2$ = {(1, 2), (1, 13), (2, 3), (2, 15), (3, 4), (6, 10), (6, 18), (7, 9), (7, 18), (10, 19)} and the corresponding coefficients (2, 2, 2, 2, 2, 1, 1, 1, 1, 1). Note that the weak heredity condition holds here.

Three additional examples, Examples 7 to 9, are listed in the Supplementary Material due to the page limit. In particular, the compound symmetry (CS) correlation is considered in Examples 7 and 9. The numerical results are summarized in the following Tables 1–6 and Tables S1–S3 in the Supplementary Material.

We first summarize the results for Examples 1–5, where the strong heredity condition holds. All the methods perform reasonably well in most of the settings, including the high dimensional cases with $p = 5, 000$ and $p = 10, 000$, as long as the noise level is not too high. Overall speaking, the iFORM is the best among all the methods in terms of both model selection and prediction performance. The iFORM method has the smallest MSE, the largest out-of-the sample R^2, and the highest exact coverage probability for main effects and interactions. When $\sigma = 2$, the iFORM's performance is quite close to the ORACL procedure. The performance of iFORT is sensitive to the dimensionality and noise level. In particular, when p is large and the noise level is high, it may miss some important main effects in Stage 1, although the result may be improved by using less aggressive selection criteria such as AIC and standard BIC. On the other hand, iFORM consistently gives higher coverage of important main effects and interactions than iFORT, which supports our motivation for the dynamic selection procedure. The FS2 has the worst performance, and it fails to run when p is 5000 or larger. Both iMART1 and iMART2 are reasonably fast and perform well, sometimes quite competitive in prediction. But when the covariance structure is complex, their performance is not very good. This can be seen in Example 5, and Examples 7 and 8 in the Supplementary Material.

In Example 6, the weak heredity condition holds, and therefore we expect that the iFOR under the weak heredity constraint should perform better than those under the strong heredity. The results in Table 6 confirm this pattern: iFORM-w (or iFORT-w) gives better performance than iFORM (or iFORT) in terms of both model selection and prediction accuracy. Since the strong heredity methods make an incorrect model structure assumption, they suffer by missing some important interactions. For example, if $\sigma = 2$, iFORM-w is the only method showing a high exact selection probability (91%) for important interactions.

Finally, we illustrate the quality of the solution path by the *hit-rate* plot. In each plot, the x-axis denotes the solution path steps $\{1, 2, \cdots, S\}$, and the y-axis represents the “hit rate” which is defined as the percentage of important terms recovered up to step s. Denote the true model size by d_0. The ideal hit plot (given by ORACL) should show a linearly increasing trend with slope $1/d_0$ within the first d_0 steps and then stays at 1 afterwards. For the graph clarity, we only draw the hit rates for the strong heredity methods. Figure 1 plots the hit-rates for Examples 1 and 2 with the moderate $p = 500$. Here $d_0 = 8$, so we choose $S = 20$. Based on Figure 1, the iFORM has the highest hit rate among all, very close to the oracle. For $\sigma = 2$ and 3, its hit rate is more than 95% after 20 steps; for the more difficult case $\sigma = 4$, iFORM still achieves approximately 90% hit rate. The iFORT is slightly worse than iFORM, with rates 90%, 80%, 70% respectively for $\sigma = 2, 3, 4$. The FS2 has the lowest hit rate, only 20% when $\sigma = 4$. Figure 2 plots the hit rates for the large p. Since $d_0 = 20$, we choose $S = 40$. The FS2 is not shown in Figure 2, because it fails to run. Again, iFORM has the highest hit rate among all (except the oracle). The iFORT is slightly worse, about 80% hit rate in most cases.

The following table summarizes the average computation time (seconds per run) for each procedure. The machine we used equips Intel Core (TM) i7-2600 CPU @ 3.40GHZ with 4.00 GB ram. Since the time difference is small for varying σ, we only present the results for $\sigma = 2$. When p is moderately large, the FS2 is slowest, taking 16.40 seconds in average for Example 1. The iFORT and iFORM are the fastest, taking 0.04 and 0.08 seconds in Example 1, which is more than 100 times faster than FS2. The weak heredity methods are slower than their strong heredity counterparts. When p is large, the FS2 fails to run, while the iFOR procedures are still amazingly fast. When $p = 5000$, it takes 11.39 (and 16.06) seconds for iFORT (and iFORM). When $p = 10000$, it takes 22.13 (and 29.17) seconds for iFORT (and iFORM). The weak heredity methods now take significantly more time. Overall, the iFORM appears the most promising in terms of both performance and speed.

6 Real Data Analysis

We analyze two real data sets, the inbred mouse microarray gene expression dataset (Lan et al. 2006) and the supermarket data (Wang, 2012). The inbred mouse microarray data set contains 60 mouse arrays, with 31 from female mice and 29 from male mice, respectively. Each array measures the expression values of 22,690 genes. The response is a continuous phenotypic variable measured by real-time RT-PCR, stearoyl-CoA desaturase 1 (SCD1). The supermarket dataset collects daily sale information of a major supermarket located in northern China, with $n = 464$ and $p = 6398$. The response Y is the number of customers per day, and the predictors $\mathbf{X}$ are sale volumes of various products. The supermarket manager is interested in the relationship between the number of customers and the sale volume of certain products. For convenience, the response and all predictors are centered to zero and standardized to have a unit variance prior to the analysis.

The proposed methods are applied to both datasets. To assess the prediction performance of the procedures, we randomly sample n_1 observations to form the training set, and use the remaining $n - n_1$ observations as the test data to compute the out-of-sample R^2 for the final model. We use $n_1 = 50$ in the inbred mouse data analysis and use $n_1 = 400$ for the supermarket data analysis. The results are summarized in Table 7. It is observed that the iFOR methods give similar performance for both data sets.

7 Discussion

In this article, we tackle the important problem of interaction selection for ultra-high dimensional data. The task is both computationally and theoretically challenging. We propose a new class of procedures, called iFOR, and study their numerical and theoretical properties. One major advantages of the proposed methods are their computation feasibility. The code is simple and fast. Theoretically we show that the iFOR can discover all relevant interactions consistently, even if the dimension increases exponentially fast with the sample size. Our numerical examples suggest that the new methods, especially iFORM, give promising performance for ultra-high dimensional data.

We use the extended BIC (Chen and Chen 2008) to select a final model from the solution path in this work. Since the motivation of the extended BIC is to control FDR, it tends to be conservative in real data analysis. It would be interesting to study the performance of other selection criteria such as AIC and cross validation for iFOR methods in the future. Other works of interest include the generalization of the iFOR to other loss functions in GLM or nonparametric regression, and how to improve computational efficiency of penalized methods with the iFOR ideas.

In practice, higher-order interactions are useful to uncover multi-way relationships among predictors for complex problems where two-way interactions are not sufficient. The proposed methods can be readily extended to selecting higher-order interactions, by including higher-order products of predictors in the candidate set. No essential change is needed in the computational algorithm, except that the enlarged candidate set will demand extra time. When considering higher-order interaction models, one should tune the model properly to avoid the over-fitting. The interpretation of higher-order interactions should be cautious as well. The topic is worth a full investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are partially supported by NSF Grants DMS-1309507 (Hao and Zhang), DMS-1347844 (Zhang), NIH Grants NIH/NCI R01 CA-085848 (Zhang) and P01 CA142538 (Zhang), AMS-Simons Travel Grants (Hao). The authors are grateful to Dr. Han Xiao and to the editors, associate editor, and four referees for their helpful comments and suggestions.

References

Bar W, Dittrich F. Useful formula for moment computation of normal random variables with nonzero means. Automatic Control, IEEE Transactions on. 1971; 16:263–265.

Bickel PJ, Levina E. Regularized estimation of large covariance matrices. Ann Statist. 2008; 36:199–227.

Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics. 2007; 35:2313–2351.

Chen J, Chen Z. Extended bayesian information criteria for model selection with large model spaces. Biometrika. 2008; 95:759–771.

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

Chipman H, Hamada M, Wu CFJ. A bayesian variable-selection approach for analyzing designed experiments with complex aliasing. Technometrics. 1997; 39:372–381.

Choi NH, Li W, Zhu J. Variable selection with the strong heredity constraint and its oracle property. Journal of the American Statistical Association. 2010; 105:354–364.

Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics. 2009; 10:392–404.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics. 2004; 32:407–451.

Evans DM, Marchini J, Morris AP, Cardon LR. Two-stage two-locus models in genome-wide association. PLoS Genet. 2006; 2:e157. [PubMed: 17002500]

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

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

Fan J, Lv J. Nonconcave penalized likelihood with np-dimensionality. Information Theory, IEEE Transactions on. 2011; 57:5467–5484.

Fan J, Peng H. On non-concave penalized likelihood with diverging number of parameters. Annals of Statistics. 2004; 32:928–961.

Fan J, Song R. Sure independence screening in generalized linear models with np-dimensionality. Annals of Statistics. 2010; 38:3567–3604.

Hamada M, Wu CFJ. Analysis of designed experiments with complex aliasing. Journal of Quality Technology. 1992; 24:130–137.

Kooperberg C, LeBlanc M. Increasing the power of identifying Gnee × Gnee interactions in Genome-Wdie Aassociation Studies. Genetic Epidemiology. 2008; 32:255–263. [PubMed: 18200600]

Laub, A. Matrix analysis for scientists and engineers. Society for Industrial and Applied Mathematics; 2005.

Manolio TA, Collins FS. Genes, environment, health, and disease: Facing up to complexity. Hum Hered. 2007:63–66. [PubMed: 17283435]

McCullagh P. What is a statistical model? The Annals of Statistics. 2002; 30:1225–1267.

McCullagh, P.; Nelder, J. Monographs on Statistics and Applied Probability. Chapman and Hall; 1989. Generalized Linear Models.

Nelder JA. A reformulation of linear models. Journal of the Royal Statistical Society. Series A (General). 1977; 140:48–77.

Nelder JA. The statistics of linear models: back to basics. Statistics and Computing. 1994; 5:i–i. 10.1007/BF00143933

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

Turlach B. Discussopn of “least angle regression”. The Annals of Statistics. 2004; 32:481–490.

van der Vaart, A.; Wellner, J. Springer series in statistics. Springer; 1996. Weak convergence and empirical processes.

Wang H. Forward regression for ultra-high dimensional variable screening. Journal of the American Statistical Association. 2009; 104:1512–1524.

Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. Screen and clean: A tool for identifying interactions in genome-wide association studies. Genetic Epidemiology. 2010; 34:275–285. [PubMed: 20088021]

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009; 25:714–721. [PubMed: 19176549]

Yuan M, Joseph V, Lin Y. An efficient variable selection approach for analyzing designed experiments. Technometrics. 2007; 49:430–439.

Yuan M, Joseph VR, Zou H. Structured variable selection and estimation. Annals of Applied Statistics. 2009; 3:1738.

Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics. 2010; 38:894–942.

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

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

Zhao P, Yu B. On model selection consistency of lasso. J Mach Learn Res. 2006; 7:2541–2563.

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67:301–320.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics. 2008; 36:1509–1533.

Appendix A. Total Covariance Matrix

In this section, we work on the total covariance matrix Σ and show it is determined by the covariance matrix $\Sigma^{(1)}$ of main effects under the Gaussian assumption (C1).

Let us temporarily ignore the index labeling the order of observations, and denote by X_j for $1 \le j \le p$ the main effects and $Z_{jk} = X_j X_k - \mathrm{E}(X_j X_k)$ for $(j, k) \in \mathcal{P}_2$ the interactions. Let $\Sigma^{(1)} = (\sigma_{ij})$ denote covariance matrix of the main effects $X_1, \ldots, X_p$. The first two lemmas help us to characterize the total covariance matrix Σ.

Lemma 1

Under the normality condition (C1), for $\forall j, k, \ell$, $\mathrm{cov}(X_j, Z_{k\ell}) = 0$ which implies

$$\sum = \begin{pmatrix} \sum^{(1)} & 0 \\ 0 & \sum^{(2)} \end{pmatrix}.$$

Proof

$\mathrm{cov}(X_j, Z_{k\ell}) = \mathrm{cov}(X_j, X_k X_\ell) = \mathrm{E}(X_j X_k X_\ell) - \mathrm{E}(X_j)\mathrm{E}(X_k X_\ell) = 0$. The conclusion still holds if the joint density of $X_1,\ldots, X_p$ is symmetric with respect to the original point $\mathbf{0}$.

Lemma 2

Under the normality condition (C1),

$$\mathrm{cov}(Z_{ij}, Z_{k\ell}) = \mathrm{cov}(X_i X_j, X_k X_\ell) = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}. \quad (7.1)$$

Proof

This lemma follows directly from the following useful formula (Bar & Dittrich, 1971)

$$\mathrm{E}(X_iX_jX_kX_\ell)=\mathrm{E}(X_iX_j)\mathrm{E}(X_kX_\ell)+\mathrm{E}(X_iX_k)\mathrm{E}(X_jX_\ell)+\mathrm{E}(X_iX_\ell)\mathrm{E}(X_jX_k)-2\mathrm{E}(X_i)\mathrm{E}(X_j)\mathrm{E}(X_k)\mathrm{E}(X_\ell).$$

Let $A = (A_{ij})$ be an $N \times N$ matrix. In linear algebra, a $K \times K$ submatrix is called a principal submatrix if it is of the form $A_{\mathcal{I}} = (A_{\ell_i\ell_j})$ where $\mathcal{I}$ is an index set $\mathcal{I} = \{1 \le \ell_1 < \cdots < \ell_K \le N\}$. Here with slight abuse of this conception, we allow arbitrary order for the index set $\mathcal{I}$. For example, let $\mathcal{I} = \{2, 1\}$ and

$$A_{\mathscr{I}}=\begin{pmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{pmatrix}$$

is still called a principal submatrix in this paper.

Based on the formula (7.1), we can decompose $\Sigma^{(2)}$ to a sum $\sum_1^{(2)}+\sum_2^{(2)}$. In fact, we have

Lemma 3

Both $\sum_1^{(2)}$ and $\sum_2^{(2)}$ are principal submatrices of $\Sigma^{(1)} \otimes \Sigma^{(1)}$.

Proof

The Kronecker product (Laub, 2005) $\Sigma^{(1)} \otimes \Sigma^{(1)}$ is a $p^2 \times p^2$ matrix whose rows and columns are both indexed by the set $\mathcal{P}_1 \times \mathcal{P}_1$. The entry corresponding to the index $(ij, k\ell)$ is $\sigma_{ij}\sigma_{k\ell}$. By formula (7.1), both $\sum_1^{(2)}$ and $\sum_2^{(2)}$ are $\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}$ principal submatrices of $\Sigma^{(1)} \otimes \Sigma^{(1)}$.

Lemma 4

Under the conditions (C1) and (C2a), we have

$$2\tau_{\min}<\lambda_{\min}\left(\sum\right) \le \lambda_{\max}\left(\sum\right)<\tau_{\max}/2. \quad (7.2)$$

Proof

By Laub (2005) Theorem 13.12, the eigenvalues of $\Sigma^{(1)} \otimes \Sigma^{(1)}$ are $\lambda_i\lambda_j$, $1 \le i, j \le p$, if the eigenvalues of $\Sigma^{(1)}$ are $\lambda_1, \ldots, \lambda_p$. Therefore, under condition (C2a), we have

$$\tau_{\min}>\lambda_{\min}\left(\sum^{(1)} \otimes \sum^{(1)}\right) \le \lambda_{\max}\left(\sum^{(1)} \otimes \sum^{(1)}\right)<\tau_{\max}/4.$$

By Lemma 3, the eigenvalues of $\sum_1^{(2)}$ and $\sum_2^{(2)}$ are also bounded by $\tau_{\min}$ and $\tau_{\max}/4$, so

$$2\tau_{\min}<\lambda_{\min}(\sum^{(2)}) \le \lambda_{\max}(\sum^{(2)})<\tau_{\max}/2.$$

It is straightforward to get (7.2).

Appendix B. A Bernstein Inequality and Its Application

In this section, we study a Bernstein-type inequality and its applications in bounding the eigenvalues of submatrices of sample covariance matrix $\hat{\Sigma}$, which is crucial in the proofs of theorems. For any index set $\mathcal{M}$, $\hat{\Sigma}_{\mathcal{M}}$ denotes the principal submatrix corresponding to $\mathcal{M}$.

Lemma 5

Let $W_1, \ldots, W_n$ be independent random variables with mean zero and variances bounded by $\sigma^2 \ge 1$. Assume for some $0 < \alpha < 1$,

$$\mathrm{E}\left(|W_i|^{3(1-\alpha)}e^{t|W_i|^{\alpha}}\right) \le A, \quad \textit{for all} \quad 1 \le i \le n, \quad 0 \le t \le T. \quad (7.3)$$

Then for $x>(\frac{2A}{T^2})^{\frac{1}{1-\alpha}}$,

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}W_i\right| \ge x\right) \le 2\exp\left\{-\frac{x^2}{2(n\sigma^2+x^{2-\alpha}/T)}\right\}+\sum_{i=1}^{n}\mathbf{P}(|W_i| \ge x). \quad (7.4)$$

Proof

Let $W_i^*=W_i \cdot I_{(-\infty,x]}(W_i)$. Then

$$\mathbf{P}\left(\sum_{i=1}^{n}W_i \ge x\right) \le \mathbf{P}\left(\sum_{i=1}^{n}W_i^* \ge x\right)+\sum_{i=1}^{n}\mathbf{P}(W_i \ge x). \quad (7.5)$$

For $W_i^* \ge 0$, we have

$$e^{tW_i^*} \le 1+tW_i^*+\frac{t^2}{2}W_i^{*2}+\sum_{k=3}^{\infty}\frac{t^k}{k!}|W_i|^{k\alpha+3(1-\alpha)}x^{(k-3)(1-\alpha)}. \quad (7.6)$$

Note that (7.6) is true also for $W_i^*<0$ because of the monotonicity of function $f(u) = e^u - 1 - u - u^2/2$.

It is easy to get $\mathrm{E}|W_i|^{k\alpha+3(1-\alpha)} \le \frac{k!A}{T^k}$ from (7.3). Moreover, we have $\mathrm{E}(W_i^*) \le 0,\ \mathrm{Var}(W_i^*) \le \sigma^2$ from definition. Taking expectation of (7.6),

$$\begin{aligned}\mathrm{E}(e^{tW_i^*}) &\le 1+\frac{t^2\sigma^2}{2}+\sum_{k=3}^{\infty}\frac{2A}{T^2x^{1-\alpha}}\frac{1}{2}\left(\frac{x^{1-\alpha}}{T}\right)^{k-2}t^k\\ &\le 1+\frac{t^2\sigma^2}{2}+\frac{t^2}{2}\sum_{k=3}^{\infty}\left(\frac{tx^{1-\alpha}}{T}\right)^{k-2}\\ &\le 1+\frac{t^2\sigma^2}{2(1-tx^{1-\alpha}/T)},\end{aligned} \qquad (7.7)$$

when $|tx^{1-\alpha}/T| < 1$.

Let $t=\frac{x}{n\sigma^2+x^{2-\alpha}/T}$. By Markov inequality

$$\begin{aligned}\mathbf{P}\left(\sum_{i=1}^{n}W_i^* \ge x\right) &\le e^{-tx}\mathrm{E}(e^{t\sum_{i=1}^{n}W_i^*})\\ &\le e^{-tx}\prod_{i=1}^{n}\mathrm{E}(e^{tW_i^*})\\ &\le e^{-tx}\left(1+\frac{t^2\sigma^2}{2(1-tx^{1-\alpha}/T)}\right)^n\\ &\le \exp\left\{-\frac{x^2}{n\sigma^2+x^{2-\alpha}/T}\right\}\left(1+\frac{1}{2n}\frac{x^2}{n\sigma^2+x^{2-\alpha}/T)}\right)^n\\ &\le \exp\left\{-\frac{x^2}{2(n\sigma^2+x^{2-\alpha}/T)}\right\}\end{aligned}$$

Therefore,

$$\mathbf{P}\left(\sum_{i=1}^{n}W_i \ge x\right) \le \exp\left\{-\frac{x^2}{2(n\sigma^2+x^{2-\alpha}/T)}\right\}+\sum_{i=1}^{n}\mathbf{P}(W_i \ge x).$$

Apply the same technique to $-W_i^*$ and combine the results, we can get (7.4).

The following is the Lemma 1 in Wang (2009), which is useful in the proof of Theorem 1.

Lemma 6

Under condition (C1) and (C2), for $m=o(n^{\frac{1}{3}-\frac{1}{3}\xi})$, $\mathcal{M} \subset \mathcal{P}_1$,

$$\mathbf{P}\left(\tau_{\min} \le \min_{|\mathcal{M}|\le m}\lambda_{\min}(\hat{\sum}_{\mathcal{M}}) \le \max_{|\mathcal{M}|\le m}\lambda_{\max}(\hat{\sum}_{\mathcal{M}}) \le \tau_{\max}\right) \to 1. \qquad (7.8)$$

Furthermore, under condition (C4), (7.8) holds for $m=O(n^{2\xi_0+4\xi_{\min}})=o(n^{\frac{1}{3}-\frac{1}{3}\xi})$.

Lemma 7

Let $W_1, \ldots, W_n$ be independent random variables with zero mean such that $E(e^{T_0|W_i|^{\alpha}}) \le A_0$ for constants $T_0 > 0$, $A_0 > 0$ and $0 < \alpha < 1$. Then, for a sequence $a_n \to \infty$ with $a_n = o(n^{\frac{\alpha}{2(2-\alpha)}})$, there exist constants c_1, c_2, such that

$$\mathbf{P}\left(|W_1 + \cdots + W_n| > \sqrt{n}a_n\right) \le c_1 \exp(-c_2 a_n^2). \quad (7.9)$$

Proof

The condition $E(e^{T_0|W_i|^{\alpha}}) \le A_0$ implies $\mathrm{Var}(W_i) \le \sigma^2$, $E(|W_i|^2 e^{T|W_i|^{\alpha}}) \le A$ and $E(|W_i|^{3(1-\alpha)} e^{T|W^i|^{\alpha}}) \le A$ for some constants σ^2, T and A. By Lemma 5, we have

$$\mathbf{P}\left(\left|\sum_{i=1}^{n} W_i\right| \ge x\right) \le 2\exp\left\{-\frac{x^2}{2(n\sigma^2 + x^{2-\alpha}/T)}\right\} + \sum_{i=1}^{n} \mathbf{P}(|W_i| \ge x).$$

Let $x = \sqrt{n}a_n$. Then

$$\exp\left\{-\frac{x^2}{2(n\sigma^2 + x^{2-\alpha}/T)}\right\} = \exp\left\{-\frac{na_n^2}{2(n\sigma^2 + n^{\frac{2-\alpha}{2}} a_n^{2-\alpha}/T)}\right\} = \exp\left\{-\frac{a_n^2}{2\sigma^2 + o(1)}\right\}.$$

On the other hand, by Markov inequality

$$\mathbf{P}(|W_i| \ge x) = \mathbf{P}(W_i^2 e^{T|W_i|^{\alpha}} \le x^2 e^{Tx^{\alpha}}) \le Ax^{-2}\exp\{-Tx^{\alpha}\} \le \frac{A}{na_n^2}\exp\{-Ta_n^2/o(1)\}.$$

Hence, $\sum_{i=1}^{n} \mathbf{P}(|W_i| \ge x) \le \frac{A}{a_n^2}\exp\{-Ta_n^2/o(1)\}$. And (7.9) is easily obtained.

Remark 1

We are interested in the case that $W_i = X_{ij}X_{ik}X_{i\ell}$, where X_{ij}, X_{ik}, $X_{i\ell}$ are joint normal and marginally standard normal. It is easy to see that W_i satisfies $E(e^{\frac{1}{4}|W_i|^{\frac{2}{3}}}) \le \sqrt{2}$ and $\mathrm{Var}(W_i) \le 30$. Therefore, (7.9) holds for $c_1 = 3$, $c_2 = 1/61$ when n is sufficiently large.

In order to show Theorem 2, we have to obtain an analogue of Lemma 6 for arbitrary submodel $\mathcal{M}$. We start from a generalization of Lemma A3 in Bickel & Levina (2008).

Lemma 8

Let $W_1, \ldots, W_n$ be independent random variables with zero mean such that $E(e^{T_0|W_i|^{\alpha}}) \le A_0$ for constants $T_0 > 0$, $A_0 > 0$ and $0 < \alpha \le 1$. Then there exist constants c_3, c_4, for $0 < \varepsilon \le 1$

NIH-PA Author Manuscript
NIH-PA Author Manuscript

$$\mathbf{P}\left(|W_1+\cdots+W_n|>n\varepsilon\right)\leq c_3\exp(-c_4n^{\alpha}\varepsilon^2). \quad (7.10)$$

Proof

The condition E($e^{T0|W_i|\alpha}$</sup>) $\leq A_0$ implies Var(W_i) $\leq \sigma^2$, E($|W_i|^2\, e^{T|W_i|^{\alpha}}$) $\leq A$ and E($|W_i|^{3(1-\alpha)}\, e^{T|W_i|^{\alpha}}$) $\leq A$ for some constants σ^2, T and A. When $\alpha < 1$, by Lemma 5,

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}W_i\right|\geq x\right)\leq 2\exp\left\{-\frac{x^2}{2(n\sigma^2+x^{2-\alpha}/T)}\right\}+\sum_{i=1}^{n}\mathbf{P}(|W_i|\ \geq x)$$

Let $x = n\varepsilon$. Then

$$\begin{aligned}\exp\left\{-\frac{x^2}{2(n\sigma^2+x^{2-\alpha}/T)}\right\}&=\exp\left\{-\frac{n^2\varepsilon^2}{2(n\sigma^2+n^{2-\alpha}\varepsilon^{2-\alpha}/T)}\right\}\\&=\exp\left\{-\frac{n^{\alpha}\varepsilon^2}{2n^{\alpha-1}\sigma^2+2\varepsilon^{2-\alpha}/T}\right\}\\&\leq\exp\left\{-\frac{n^{\alpha}\varepsilon^2}{o(1)+2/T}\right\}.\end{aligned}$$

On the other hand, by Markov inequality

$$\mathbf{P}(|W_i|\ \geq x)=\mathbf{P}(W_i^2e^{T|W_i|^{\alpha}}\leq x^2e^{Tx^{\alpha}})\leq Ax^{-2}\exp\{-Tx^{\alpha}\}\leq\frac{A}{n^2\varepsilon^2}\{-Tn^{\alpha}\varepsilon^{\alpha}\}.$$

Hence, $\sum_{i=1}^{n}\mathbf{P}(|W_i|\ \geq x)\leq\frac{A}{n\varepsilon^2}\exp\{-\frac{1}{2}Tn^{\alpha}\varepsilon^{\alpha}\}\exp\{-\frac{1}{2}Tn^{\alpha}\varepsilon^{\alpha}\}\leq o(1)\exp\{-\frac{1}{2}Tn^{\alpha}\varepsilon^2\}$. And (7.10) is easily obtained.

When $\alpha = 1$, E($e^{T0|W_i|}$) $\leq A_0$ implies $\sum_{k=0}^{\infty}\frac{1}{k!}T_0^k\mathrm{E}(|W_i|^k)\leq A_0$. So $\mathrm{E}(|W_i|^k)\leq\frac{1}{2}k!\left(\frac{1}{T_0}\right)^{k-2}\frac{2A_0}{T_0^2}$ for $k \geq 2$. By Bernstein's inequality, Lemma 2.2.11 in van der Vaart & Wellner (1996), we have

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}W_i\right|\geq n\varepsilon\right)\leq 2\exp\{-\frac{n^2\varepsilon^2}{2(2nA_0/T_0^2+n\varepsilon/T_0)}\}\leq 2\exp\{-\frac{n\varepsilon^2}{4A_0/T_0^2+2/T_0}\}.$$

Lemma 9

Under condition (C1) and (C2), for $0 < \varepsilon < 1$, we have

$$\mathbf{P}\left(\left|\sum_{s=1}^{n}(X_{si}X_{sj}-\sigma_{ij})\right|\geq n\varepsilon\right)\leq C_1\exp(-C_2n\varepsilon^2) \quad (7.11)$$

$$\mathbf{P}\left(\left|\sum_{s=1}^{n}(X_{si}X_{sj}X_{sk}-0)\right| \geq n\varepsilon\right) \leq C_3\exp(-C_4 n^{\frac{2}{3}}\varepsilon^2) \quad (7.12)$$

$$\mathbf{P}\left(\left|\sum_{s=1}^{n}(X_{si}X_{sj}X_{sk}X_{s\ell}-\sigma_{ij}\sigma_{k\ell}-\sigma_{ik}\sigma_{j\ell}-\sigma_{i\ell}\sigma_{jk})\right| \geq n\varepsilon\right) \leq C_5\exp(-C_6 n^{\frac{1}{2}}\varepsilon^2) \quad (7.13)$$

where $C_1, \ldots, C_6$ are constants.

Proof

We show the last inequality here. The first two are similar. Let $W_s = X_{si}X_{sj}X_{sk}X_{s\ell} - \sigma_{ij}\sigma_{k\ell} - \sigma_{ik}\sigma_{j\ell} - \Sigma_{i\ell}\,\sigma_{jk}$.

$$\begin{aligned}
\mathrm{E}\left(e^{\frac{1}{4}|W_s|^{\frac{1}{2}}}\right) &= \mathrm{E}\left(e^{\frac{1}{4}|X_{si}X_{sj}X_{sk}X_{s\ell}-\sigma_{ij}\sigma_{k\ell}-\sigma_{ik}\sigma_{j\ell}-\sigma_{i\ell}\sigma_{jk}|^{\frac{1}{2}}}\right)\\
(\because (a+b)^{\frac{1}{2}} \leq a^{\frac{1}{2}}+b^{\frac{1}{2}}) &\leq \mathrm{E}\left(e^{\frac{1}{4}|X_{si}X_{sj}X_{sk}X_{s\ell}|^{\frac{1}{2}}+\frac{1}{4}|\sigma_{ij}\sigma_{k\ell}+\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk}|^{\frac{1}{2}}}\right)\\
(\because |\sigma_{ij}\sigma_{k\ell}+\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk}| \leq 3) &\leq e^{\frac{\sqrt{3}}{4}}\mathrm{E}\left(e^{\frac{1}{4}|X_{si}X_{sj}X_{sk}X_{s\ell}|^{\frac{1}{2}}}\right)\\
(\because abcd \leq \tfrac{a^4+b^4+c^4+d^4}{4}) &\leq e^{\frac{\sqrt{3}}{4}}\mathrm{E}\left(e^{\frac{1}{4}\frac{X_{si}^2+X_{sj}^2+X_{sk}^2+X_{s\ell}^2}{4}}\right)\\
(\text{again } abcd \leq \tfrac{a^4+b^4+c^4+d^4}{4}) &\leq e^{\frac{\sqrt{3}}{4}}\mathrm{E}\left([e^{\frac{X_{si}^2}{4}}+e^{\frac{X_{sj}^2}{4}}+e^{\frac{X_{sk}^2}{4}}+e^{\frac{X_{s\ell}^2}{4}}]/4\right)\\
&= \sqrt{2}e^{\frac{\sqrt{3}}{4}}
\end{aligned}$$

The inequality follows directly from the last lemma.

Lemma 10

Under condition (C1) and (C2a), for $m=o(n^{\frac{1}{6}-\frac{1}{3}\xi})$,

$$\mathbf{P}\left(\tau_{\min} \leq \min_{|\mathcal{M}|\leq m}\lambda_{\min}(\hat{\sum}_{\mathcal{M}}) \leq \max_{|\mathcal{M}|\leq m}\lambda_{\max}(\hat{\sum}_{\mathcal{M}}) \leq \tau_{\max}\right) \to 1. \quad (7.14)$$

Furthermore, under condition (C4), (7.14) holds for $m=O(n^{2\xi_0+4\xi_{\min}})=o(n^{\frac{1}{6}-\frac{1}{3}\xi})$.

Proof

The proof is similar to Lemma 1 in Wang (2009), where the inequality (7.11) plays a crucial role. The inequality (7.11) implies $\mathbf{P}\left(|\hat{\sum}_{ij}^{(1)}-\sum_{ij}^{(1)}|>\varepsilon\right) \leq C_1\exp(-C_2 n\varepsilon^2)$ for $\forall 1 \leq i, j \leq p$. Since the distribution of interactions have heavier tails, we have

$$\mathbf{P}\left(|\hat{\sum}_{\kappa\gamma}-\hat{\sum}_{\kappa\gamma}|>\varepsilon\right)\leq C_7\exp(-C_8 n^{\frac{1}{2}}\varepsilon^2), \quad (7.15)$$

for $\forall\kappa,\ \gamma\in\mathcal{P}_1\cup\mathcal{P}_2$. For example, if $\kappa=(i,j),\ \gamma=(k,\ell)\in\mathcal{P}_2$,

$$\begin{aligned}|\hat{\sum}_{\kappa\gamma}-\sum_{\kappa\gamma}|&=|\tfrac{1}{n}\sum_{s=1}^{n}(X_{si}X_{sj}-\hat{\sum}_{ij})(X_{sk}X_{s\ell}-\hat{\sum}_{k\ell})-(\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})|\\&=|\tfrac{1}{n}\sum_{s=1}^{n}X_{si}X_{sj}X_{sk}X_{s\ell}-\hat{\sum}_{ij}^{(1)}\hat{\sum}_{k\ell}^{(1)}-(\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})|\\&\leq|\tfrac{1}{n}\sum_{s=1}^{n}X_{si}X_{sj}X_{sk}X_{s\ell}-(\sigma_{ij}\sigma_{k\ell}+\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})|+|\hat{\sum}_{ij}^{(1)}\hat{\sum}_{k\ell}^{(1)}-\sigma_{ij}\sigma_{k\ell}|\\&\leq|\tfrac{1}{n}\sum_{s=1}^{n}X_{si}X_{sj}X_{sk}X_{s\ell}-(\sigma_{ij}\sigma_{k\ell}+\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})|+|\hat{\sum}_{ij}^{(1)}(\hat{\sum}_{k\ell}^{(1)}-\sigma_{k\ell})|+|(\hat{\sum}_{ij}^{(1)}-\sigma_{ij})\sigma_{k\ell}\\&\leq|\tfrac{1}{n}\sum_{s=1}^{n}X_{si}X_{sj}X_{sk}X_{s\ell}-(\sigma_{ij}\sigma_{k\ell}+\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})|+|\hat{\sum}_{k\ell}^{(1)}-\sigma_{k\ell}|+|\hat{\sum}_{ij}^{(1)}-\sigma_{ij}|\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{P}\left(|\hat{\sum}_{\kappa\gamma}-\hat{\sum}_{\kappa\gamma}|>\varepsilon\right)&\leq\mathbf{P}\left(|\tfrac{1}{n}\sum_{s=1}^{n}X_{si}X_{sj}X_{sk}X_{s\ell}-(\sigma_{ij}\sigma_{k\ell}+\sigma_{ik}\sigma_{j\ell}+\sigma_{i\ell}\sigma_{jk})|>\tfrac{\varepsilon}{3}\right)+\mathbf{P}\left(|\hat{\sum}_{k\ell}^{(1)}-\sigma_{k\ell}|>\tfrac{\varepsilon}{3}\right)+\mathbf{P}\left(|\hat{\sum}_{ij}^{(1)}-\sigma_{ij}|>\tfrac{\varepsilon}{3}\right)\\&\leq C_5\exp(-C_6 n^{\frac{1}{2}}(\varepsilon/3)^2)+2C_1\exp(-C_2 n(\varepsilon/3)^2)\\&\leq C_7\exp(-C_8 n^{\frac{1}{2}}\varepsilon^2).\end{aligned}$$

Let $\mathbf{v}=(v_1,\ldots,v_p,v_{11},\ldots,v_{pp})^\top$ be a $p+p(p+1)/2$ dimensional vector and $\mathbf{v}_{\mathcal{M}}$ be the subvector corresponding to index set $\mathcal{M}\subset\mathcal{P}_1\cup\mathcal{P}_2=\mathcal{F}$. Recall $\Sigma_{\mathcal{M}}$ is the principle submatrix corresponding to $\mathcal{M}$. By Lemma 4, we have

$$2\tau_{\min}<\min_{\mathscr{M}\subset\mathscr{F}}\inf_{\|\mathbf{v}_{\mathscr{M}}\|=1}\mathbf{v}_{\mathscr{M}}^\top\sum_{\mathscr{M}}\mathbf{v}_{\mathscr{M}}\leq\max_{\mathscr{M}\subset\mathscr{F}}\sup_{\|\mathbf{v}_{\mathscr{M}}\|=1}\mathbf{v}_{\mathscr{M}}^\top\sum_{\mathscr{M}}\mathbf{v}_{\mathscr{M}}<\tau_{\max}/2.$$

To show (7.14), it suffices to show

$$\mathbf{P}\left(\max_{|\mathscr{M}|\leq m}\sup_{\|\mathbf{v}_{\mathscr{M}}\|=1}|\mathbf{v}_{\mathscr{M}}^\top(\hat{\sum}_{\mathscr{M}}-\sum_{\mathscr{M}})\mathbf{v}_{\mathscr{M}}|>\varepsilon\right)\to 0, \quad (7.16)$$

for arbitrarily small positive number ε. The left-hand side of (7.16) is bounded by

$$\sum_{|\mathscr{M}|\leq m}\sum_{\kappa,\gamma\in\mathscr{F}}\mathbf{P}\left(\left|\hat{\sum}_{\kappa\gamma}-\sum_{\kappa\gamma}\right|>\frac{\varepsilon}{m}\right). \quad (7.17)$$

Note that the number of possible models with sizes smaller than m is less than $(p+p(p+1)/2)^m\leq p^{2m}$ when $p\geq 3$. Applying (7.15), we can bound (7.17) further

$$(7.17) \le p^{2m}(p^2)^2 C_7 \exp(-C_8 n^{\frac{1}{2}} \varepsilon^2/m^2) \quad (7.18)$$

$$= C_7 \exp((2m+4)) \log p - C_8 n^{\frac{1}{2}} \varepsilon^2/m^2) \quad (7.19)$$

$$\le C_7 \exp(2m\nu n^{\xi}(1 - \frac{1}{2} C_8 \nu^{-1} \varepsilon^2 n^{\frac{1}{2}-\xi} m^{-3})), \quad (7.20)$$

which converges to zero when $n \to \infty$ and $m = o(n^{\frac{1}{6}-\frac{1}{3}\xi})$.

Remark 2

Beyond normality. Lemmas 6, 7, 10 play important roles in the proofs of Theorems 1 and 2. A key assumption is $E(e^{T^0|W_i|^{\alpha}}) \le A_0$ where W_i is (higher) product of predictors. It is easy to see that the condition still holds, using the argument of Lemma 9, if the marginal distributions of **X** is subGaussian. In particular, Theorem 1 is still true if condition (C1′) holds and Theorem 2 is still true if (C1′) holds and the total covariance matrix Σ has bounded eigenvalues asymptotically.

Appendix C. Proofs of Theorem 1 and 2

With slight abuse of notations, we denote by **X** the total design matrix including main and interaction effects. For any index set $\mathcal{M} \subset \mathcal{F}$, $\mathbf{X}_{(\mathcal{M})}$ is the submatrix of **X** whose columns correspond to $\mathcal{M}$; $\beta_{(\mathcal{M})}$ is the subvector of β corresponding to $\mathcal{M}$. If $\mathcal{M} = \{j\}$, we simply use $\mathbf{X}_j$ and β_j.

We first overview the general strategy (in the context of FS2) and then give proofs for theorems. The goal is to show that all important predictors in the candidate pool are selected within a number of steps, for FS2 and the first stage of iFORT. By the nature of FS, the predictors are selected sequentially, one at each step. Therefore, we divide the whole procedure into a sequence of stages, each of which consists of several steps, starting immediately after one important term is selected and ending when the next predictor is identified. If we can show that the length of each stage is less than some integer L, then after d_0L steps, all important predictors would have been selected.

Assume that stage T is the earliest stage among all that lasts longer than L steps, and $T < d_0$. Working within stage T, we omit the stage label T, and denote by $\mathcal{S}_t$ the index set of all selected predictors up to step t of stage T. Define

$$\Omega(t) = \mathrm{RSS}(\mathcal{S}_t) - \mathrm{RSS}(\mathcal{S}_{t+1}),$$

where RSS($\mathcal{S}_t$) is the residual sum of squares of **Y** regressed on the predictor space spanned by $\mathcal{S}_t$. A key step is to show that

$$n^{-1}\Omega(t) \geq 2L^{-1}(1-o(1)) \quad \text{for all} \quad 1 \leq t \leq L. \quad (7.21)$$

Therefore, we have $n^{-1}\|\boldsymbol{Y}\|^2 \geq \sum_{t=1}^{L} n^{-1}\Omega(t) \geq 2(1-o(1)) \to 2$, which contradicts with the fact Var(Y) = 1. Then we can conclude that every stage contains less than L steps.

The inequalities of type (7.21) are obtained in the following proofs, which lead to Theorems 1 and 2. We illustrate Theorem 2 first, because it is technically more straightforward.

Proof of Theorem 2

Given the regularity conditions and Lemma 10, the proof of Theorem 2 is similar to that of Theorem 1 in Wang (2009). Let $K=2\tau_{\max}\nu C_{\beta}^{2}\tau_{\min}^{-1}\nu_{\beta}^{-4}$ and $L = Kn^{\xi_0+4\xi_{\min}}$. Note that $|\mathscr{S}_t| < d_0L \leq K\nu n^{2\xi_0+4\xi_{\min}}$, so the eigenvalues of $\Sigma_{\mathscr{M}}$ can be controlled by Lemma 10. Following (B. 1) and (B.2) in Wang (2009), we have

$$\Omega(t)^{\frac{1}{2}} \geq \max_{j\in\mathscr{T}}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}\mathbf{X}_{(\mathscr{T})}\boldsymbol{\beta}_{(\mathscr{T})}\| - \max_{j\in\mathscr{T}}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}\boldsymbol{\varepsilon}\|, \quad (7.22)$$

where $\mathbf{Q}_{(\mathscr{S}_t)}=\mathbf{I}_n-\mathbf{H}_{(\mathscr{S}_t)}=\mathbf{I}_n-\mathbf{X}_{(\mathscr{S}_t)}(\mathbf{X}_{(\mathscr{S}_t)}^{\top}\mathbf{X}_{(\mathscr{S}_t)})^{-1}\mathbf{X}_{(\mathscr{S}_t)}^{\top}, \mathbf{H}_j^{(t)}=\mathbf{X}_j^{(t)}\mathbf{X}_j^{(t)\top}\|\mathbf{X}_j^{(t)}\|^{-2}$ and $\mathbf{X}_j^{(t)}=(\mathbf{I}_n-\mathbf{H}_{(\mathscr{S}_t)})\mathbf{X}_j$.

Following the procedure leading to (B.7) in Wang (2009), we have, with probability tending to 1,

$$\max_{j\in\mathscr{T}}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}\mathbf{X}_{(\mathscr{T})}\boldsymbol{\beta}_{(\mathscr{T})}\|^2 \geq \tau_{\max}^{-1}\nu^{-1}C_{\beta}^{-2}\tau_{\min}^{2}\nu_{\beta}^{4}n^{1-\xi_0-4\xi_{\min}}. \quad (7.23)$$

Similar to (B.8) in Wang (2009),

$$\max_{j\in\mathscr{T}}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}\boldsymbol{\varepsilon}\|^2 \leq \tau_{\min}^{-1}n^{-1}\max_{j\in\mathscr{T}}\max_{|\mathscr{M}|\leq m^*}(\mathbf{X}_j^{\top}\mathbf{Q}_{(\mathscr{M})}\boldsymbol{\varepsilon})^2, \quad (7.24)$$

where $m^* \leq TL \leq d_0L$. Given $\mathbf{X}$, $\mathbf{X}_j^{\top}\mathbf{Q}_{(M)}\boldsymbol{\varepsilon}$ is a normal random variable with mean 0 and variance $\|\mathbf{Q}_{(\mathscr{M})}\mathbf{X}_j\|^2 \leq \|\mathbf{X}_j\|^2$. So (7.24) is further bounded by

$$\leq \tau_{\min}^{-1}n^{-1}\max_{j\in\mathscr{T}}\|\mathbf{X}_j\|^2\max_{j\in\mathscr{T}}\max_{|\mathscr{M}|\leq m^*}\chi_1^2,$$

where χ_1^2 represents a chi-square random variable with one degree of freedom. By Lemma 10, $n^{-1}\max_{j\in\mathscr{T}}\|\mathbf{X}_j\|^2 \leq \tau_{\max}$ with probability tending to one. Moreover, the total number of combinations for $j \in \mathscr{T}$ and $|\mathscr{M}| \leq m^*$ is no more than $(p^2)^{m^*+2} = p^{2m^*+4}$. Therefore,

$$\max_{j\in\mathscr{T}}\max_{|\mathscr{M}|\le m^*}\chi_1^2\le 2(2m^*+4)\log p \\ \le 5d_0L\nu n^{\xi} \\ \le 5K\nu^2n^{\xi+2\xi_0+4\xi_{\min}}$$

with probability tending to one. Finally, we have

$$n^{-1}\Omega(t)\ge n^{-1}\left((\tau_{\max}^{-1}\nu^{-1}C_{\beta}^{-2}\tau_{\min}^2\nu_{\beta}^4n^{1-\xi_0-4\xi_{\min}})^{\frac{1}{2}}-(\tau_{\min}^{-1}\tau_{\max}5K\nu^2n^{\xi+2\xi_0+4\xi_{\min}})^{\frac{1}{2}}\right)^2 \\ \ge \tau_{\max}^{-1}\nu^{-1}C_{\beta}^{-2}\tau_{\min}^2\nu_{\beta}^4n^{-\xi_0-4\xi_{\min}}(1-2(\tau_{\max}^2\nu^3C_{\beta}^2\tau_{\min}^{-3}\nu_{\beta}^{-4}5Kn^{\xi+3\xi_0+8\xi_{\min}-1})^{\frac{1}{2}}) \\ =2L^{-1}(1-o(1)).$$

Proof of Theorem 1

Because we concentrate on only main effects in the first stage of iFORT, similar to (7.22), we have

$$\Omega(t)^{\frac{1}{2}}\ge\max_{j\in\mathscr{T}_1}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}\mathbf{X}_{(\mathscr{T}_1)}\boldsymbol{\beta}_{(\mathscr{T}_1)}\|-\max_{j\in\mathscr{T}_1}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}(\mathbf{X}_{(\mathscr{T}_2)}\boldsymbol{\beta}_{(\mathscr{T}_2)}+\boldsymbol{\varepsilon})\|. \quad (7.25)$$

The first term on the right hand side can be bounded as

$$\max_{j\in\mathscr{T}_1}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}\mathbf{X}_{(\mathscr{T}_1)}\boldsymbol{\beta}_{(\mathscr{T}_1)}\|^2\ge\tau_{\max}^{-1}\nu^{-1}C_{\beta}^{-2}\tau_{\min}^2\nu_{\beta}^4n^{1-\xi_0-4\xi_{\min}}. \quad (7.26)$$

Similar to (7.24),

$$\max_{j\in\mathscr{T}_1}\|\mathbf{H}_j^{(t)}\mathbf{Q}_{(\mathscr{S}_t)}(\mathbf{X}_{(\mathscr{T}_2)}\boldsymbol{\beta}_{(\mathscr{T}_2)}+\boldsymbol{\varepsilon})\|^2\le\tau_{\min}^{-1}n^{-1}\max_{j\in\mathscr{T}_1}\max_{|\mathscr{M}|\le m^*}(\mathbf{X}_j^{\top}\mathbf{Q}_{(\mathscr{M})}(\mathbf{X}_{(\mathscr{T}_2)}\boldsymbol{\beta}_{(\mathscr{T}_2)}+\boldsymbol{\varepsilon}))^2 \\ \le 3\tau_{\min}^{-1}n^{-1}\max_{j\in\mathscr{T}_1}\max_{|\mathscr{M}|\le m^*}\left((\mathbf{X}_j^{\top}\mathbf{X}_{(\mathscr{T}_2)}\boldsymbol{\beta}_{(\mathscr{T}_2)})^2+(\mathbf{X}_j^{\top}\mathbf{H}_{(\mathscr{M})}\mathbf{X}_{(\mathscr{T}_2)}\boldsymbol{\beta}_{(\mathscr{T}_2)})^2+(\mathbf{X}_j^{\top}\mathbf{Q}_{(\mathscr{M})}\boldsymbol{\varepsilon})^2\right) \quad (7.27)$$

where $m^*\le TL\le p_0L$.

For the first term in (7.27),

$$(\mathbf{X}_j^{\top}\mathbf{X}_{(\mathscr{T}_2)}\boldsymbol{\beta}_{(\mathscr{T}_2)})^2=(\sum_{\kappa\in\mathscr{T}_2}\mathbf{X}_j^{\top}\mathbf{X}_{\kappa}\boldsymbol{\beta}_{(\kappa)})^2\le q_0\left(\max_{\kappa\in\mathscr{T}_2}|\mathbf{X}_j^{\top}\mathbf{X}_{\kappa}|\right)^2\|\boldsymbol{\beta}_{\mathscr{T}_2}\|^2.$$

Therefore,

$$3\tau_{\min}^{-1}n^{-1}\max_{j\in\mathscr{T}_1}\max_{|\mathscr{M}|\le m^*}(\mathbf{X}_j^{\top}\mathbf{X}_{(\mathscr{T}_2)}\boldsymbol{\beta}_{(\mathscr{T}_2)})^2\le 3\tau_{\min}^{-1}n^{-1}q_0C_{\beta}\max_{j\in\mathscr{T}_1}\max_{\kappa\in\mathscr{T}_2}(\mathbf{X}_j^{\top}\mathbf{X}_{\kappa})^2. \quad (7.28)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

By Lemma 7, Remark 1 and Bonferroni inequality,

$$\mathbf{P}(\max_{j\in\mathcal{T}_1}\max_{\kappa\in\mathcal{T}_2}(\mathbf{X}_j^\top\mathbf{X}_\kappa)>\sqrt{n}20\sqrt{\log n})\le p_0q_03\exp(-400\log n/61)\le\exp(2\log\nu+2\xi_0\log n-2\log n)\to 0.$$

Thus (7.28) can be bounded by $1200\tau_{\min}^{-1}C_{\boldsymbol{\beta}}\nu n^{\xi_0}\log n$ with probability tending to 1.

For the second term,

$$\begin{aligned}(\mathbf{X}_j^\top\mathbf{H}_{(\mathcal{M})}\mathbf{X}_{(\mathcal{T}_2)}\boldsymbol{\beta}_{(\mathcal{T}_2)})^2&=(\sum_{\kappa\in\mathcal{T}_2}\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M})}(\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_{(\mathcal{M})})^{-1}\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_\kappa\boldsymbol{\beta}_{(\kappa)})^2\\&\le q_0\left(\max_{\kappa\in\mathcal{T}_2}\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M})}(\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_{(\mathcal{M})})^{-1}\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_\kappa\right)^2\|\boldsymbol{\beta}_{\mathcal{T}_2}\|^2.\end{aligned}$$

Therefore,

$$\begin{aligned}3\tau_{\min}^{-1}n^{-1}\max_{j\in\mathcal{T}_1}\max_{|\mathcal{M}|\le m^*}(\mathbf{X}_j^\top\mathbf{H}_{(\mathcal{M})}\mathbf{X}_{(\mathcal{T}_2)}\boldsymbol{\beta}_{(\mathcal{T}_2)})^2&\le 3\tau_{\min}^{-1}n^{-1}q_0C_{\boldsymbol{\beta}}\max_{j\in\mathcal{T}_1}\max_{|\mathcal{M}|\le m^*}\max_{\kappa\in\mathcal{T}_2}(\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M})}(\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_{(\mathcal{M})})^{-1}\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_\kappa)^2\\&\ge 3\tau_{\min}^{-1}n^{-1}q_0C_{\boldsymbol{\beta}}\max_{j\in\mathcal{T}_1}\max_{|\mathcal{M}|\le m^*}\max_{\kappa\in\mathcal{T}_2}\left(\|\mathbf{X}_j^T\mathbf{X}_{(\mathcal{M})}(\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_{(\mathcal{M})})^{-1}\|_\infty m^*\max_{\ell\in\mathcal{M}}|\mathbf{X}_\ell^\top\mathbf{X}_\kappa|\right)^2\\&\le 3\tau_{\min}^{-1}n^{-1}q_0C_{\boldsymbol{\beta}}m^{*2}\max_{j\in\mathcal{T}_1}\max_{|\mathcal{M}|\le m^*}\|\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M})}(\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_{(\mathcal{M})})^{-1}\|_\infty^2\max_{\kappa\in\mathcal{T}_2}\max_{\ell\in\mathcal{P}_1}\left(\mathbf{X}_\ell^\top\mathbf{X}_\kappa\right)^2\end{aligned}\quad(7.29)$$

where $\|\cdot\|_\infty$ denote the vectorized infinity norm. By Lemma 6,

$$\|\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M})}(\mathbf{X}_{(\mathcal{M})}^\top\mathbf{X}_{(\mathcal{M})})^{-1}\|_\infty\le\|\frac{\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M})}}{n}\|_2\|\left(\frac{\mathbf{X}_{(\mathcal{M})}^T\mathbf{X}_{(\mathcal{M})}}{n}\right)^{-1}\|_2\le\tau_{\max}\tau_{\min}^{-1},$$

with probability tending to one. By Lemma 7,

$$\mathbf{P}(\max_{\kappa\in\mathcal{T}_2}\max_{\ell\in\mathcal{P}_1}(\mathbf{X}_\ell^\top\mathbf{X}_\kappa)>\sqrt{n}\sqrt{100\nu n\xi})\le pq_03\exp(-\frac{200\nu n\xi}{61})\le 3\exp(\nu n^\xi+\log\nu+\xi_0\log n-\frac{100}{61}\nu n\xi)\to 0.$$

Thus, with probability tending to one, (7.29) is further bounded by

$$300\tau_{\max}^2\tau_{\min}^{-3}C_{\boldsymbol{\beta}}m^{*2}\nu^2n^{\xi_0+\xi}\le 300\tau_{\max}^2\tau_{\min}^3C_{\boldsymbol{\beta}}\nu^4K^2n^{5\xi_0+8\xi_{\min}+\xi}.\quad(7.30)$$

Following the same steps after (7.24), the third term in (7.27) can be controlled by,

$$15\tau_{\min}^{-1}\tau_{\max}K\nu^2n^{\xi+2\xi_0+4\xi_{\min}}\quad(7.31)$$

NIH-PA Author Manuscript
NIH-PA Author Manuscript

Finally, combining all results, we have

$$\begin{aligned}\Omega(t)^{\frac{1}{2}} \geq (\tau_{\max}^{-1}\nu^{-1}C_{\beta}^{-2}\tau_{\min}^{2}\nu_{\beta}^{4}n^{1-\xi_0-4\xi_{\min}})^{\frac{1}{2}} - \\ (1200\tau_{\min}^{-1}C_{\beta}\nu n^{\xi_0}\log n + 300\tau_{\max}^{2}\tau_{\min}^{-3}C_{\beta}\nu^4K^2n^{5\xi_0+8\xi_{\min}+\xi} + 15\tau_{\min}^{-1}\tau_{\max}K\nu^2n^{\xi+2\xi_0+4\xi_{\min}})^{\frac{1}{2}} \\ \geq n^{\frac{1}{2}}(\tau_{\max}^{-1}\nu^{-1}C_{\beta}^{-2}\tau_{\min}^{2}\nu_{\beta}^{4}n^{-\xi_0-4\xi_{\min}})^{\frac{1}{2}} \times (1-A_1-A_2-A_3)^{\frac{1}{2}}\end{aligned}$$

where $A_1 = 1200\tau_{\min}^{-3}\tau_{\max}C_{\beta}^{3}\nu^2\nu_{\beta}^{-4}n^{2\xi_0+4\xi_{\min}-1}\log n$, $A_2 = 300\tau_{\max}^{3}\tau_{\min}^{-5}C_{\beta}^{3}\nu_{\beta}^{-4}\nu^5K^2n^{6\xi_0+12\xi_{\min}+\xi-1}$, $A_3 = 15\tau_{\min}^{-3}\tau_{\max}^{2}K\nu^3C_{\beta}^{2}\nu_{\beta}^{-4}n^{\xi+3\xi_0+8\xi_{\min}-1}$. Therefore,

$$n^{-1}\Omega(t) \geq 2L^{-1}(1-o(1)).$$

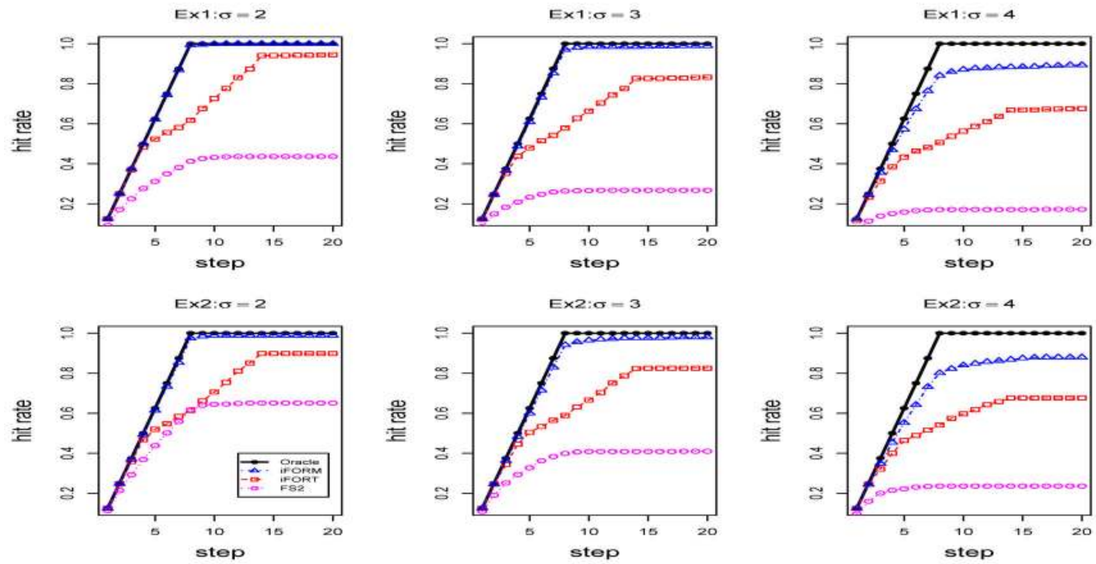

Figure 1.

The hit-rate plots for the moderate p for ORACL, iFORM, iFORT, and FS2.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

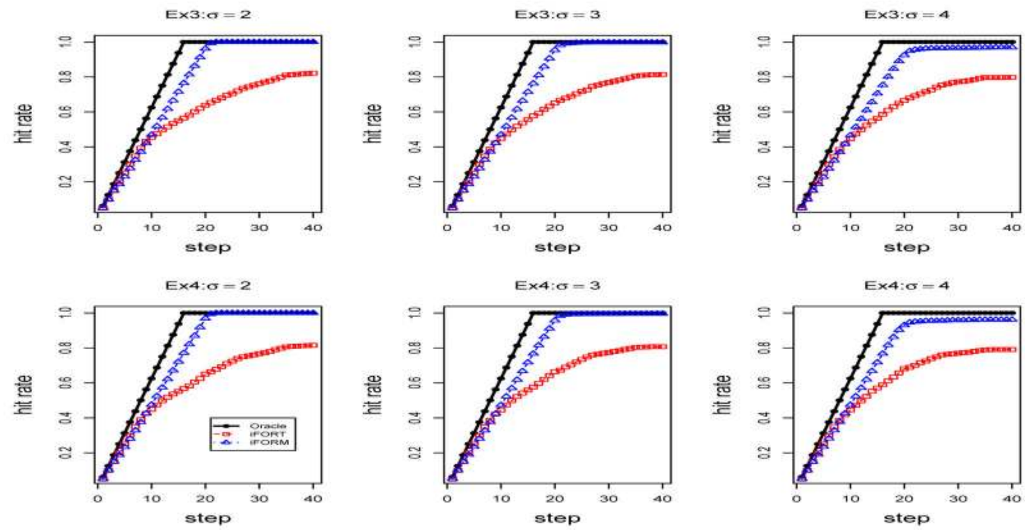

Figure 2.

The hit-rate rate plots for large *p* for ORACL, iFORM, iFORT, and FS2.

NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 1

Results of Example 1, $(n, p, p_0, q_0) = (100, 500, 4, 4)$, independent predictors.

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	**Cor0**	**Inc0**	**Ext**	**iCov**	**iCor0**	**iInc0**	**Iext**	**size**	**MSE**	**Rsq**	**sdR**
						$\sigma = 2$						
FS2	0.34	1.00	0.55	0.34	0.22	1.00	0.73	0.14	3.32	5.27	30.00	4.25
iMART1	0.71	1.00	0.07	0.42	0.71	1.00	0.17	0.29	10.02	2.39	77.21	2.07
iMART2	0.69	1.00	0.08	0.67	0.06	1.00	0.45	0.06	8.24	3.55	67.29	1.74
iFORT-w	0.84	1.00	0.06	0.82	0.77	1.00	0.17	0.60	7.48	1.63	81.99	2.12
iFORT	0.84	1.00	0.06	0.82	0.84	1.00	0.11	0.63	7.71	1.35	84.57	1.84
iFORM-w	0.98	1.00	0.01	0.98	0.90	1.00	0.08	0.60	8.13	1.05	88.24	1.32
iFORM	1.00	1.00	0.00	0.96	0.99	1.00	0.00	0.99	8.07	0.63	91.89	0.25
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	8.00	0.59	92.03	0.22
						$\sigma = 3$						
FS2	0.03	1.00	0.83	0.03	0.01	1.00	0.95	0.01	1.06	6.91	4.93	2.18
iMART1	0.59	1.00	0.12	0.30	0.53	1.00	0.29	0.13	11.20	3.48	56.91	2.98
iMART2	0.58	1.00	0.12	0.56	0.06	1.00	0.53	0.04	8.46	4.08	53.72	2.16
iFORT-w	0.63	1.00	0.17	0.62	0.22	1.00	0.61	0.17	5.22	3.86	54.64	2.45
iFORT	0.63	1.00	0.17	0.60	0.63	1.00	0.29	0.48	6.45	2.62	65.52	2.73
iFORM-w	0.69	1.00	0.16	0.69	0.24	1.00	0.59	0.13	5.39	3.72	56.16	2.54
iFORM	0.98	1.00	0.01	0.95	0.74	1.00	0.14	0.74	7.52	1.48	79.31	1.08
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	8.00	0.89	83.61	0.43
						$\sigma = 4$						
FS2	0.01	1.00	0.90	0.01	0.00	1.00	0.98	0.00	0.67	7.13	0.39	1.72
iMART1	0.49	1.00	0.15	0.28	0.34	1.00	0.44	0.05	11.63	4.49	39.12	2.73
iMART2	0.46	1.00	0.15	0.43	0.05	1.00	0.59	0.04	8.50	4.60	41.03	2.10
iFORT-w	0.41	1.00	0.33	0.41	0.02	1.00	0.90	0.01	3.36	5.17	33.99	2.15
iFORT	0.41	1.00	0.33	0.38	0.40	1.00	0.50	0.29	5.02	4.00	44.46	2.89
iFORM-w	0.22	1.00	0.50	0.22	0.03	1.00	0.90	0.03	2.53	5.67	25.92	2.28
iFORM	0.67	1.00	0.19	0.67	0.15	1.00	0.64	0.15	4.72	3.93	48.14	2.46

NIH-PA Author Manuscript
NIH-PA Author Manuscript

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	**Cor0**	**Inc0**	**Ext**	**iCov**	**iCor0**	**iInc0**	**Iext**	**size**	**MSE**	**Rsq**	**sdR**
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	8.00	1.19	73.97	0.63

Table 2

Results of Example 2, (n, p, p_0, q_0) = (100, 500, 4, 4), AR(0.5) correlation.

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	**Cor0**	**Inc0**	**Ext**	**iCov**	**iCor0**	**iInc0**	**Iext**	**size**	**MSE**	**Rsq**	**sdR**
						$\sigma = 2$						
FS2	0.55	1.00	0.31	0.55	0.43	1.00	0.52	0.24	5.14	3.76	58.43	3.77
iMART1	0.81	1.00	0.05	0.65	0.81	1.00	0.12	0.29	10.43	2.08	82.71	2.11
iMART2	0.80	1.00	0.05	0.80	0.11	1.00	0.41	0.10	7.96	3.23	75.54	1.54
iFORT-w	0.71	1.00	0.11	0.71	0.65	1.00	0.25	0.55	7.04	2.29	79.69	2.18
iFORT	0.71	1.00	0.11	0.71	0.71	1.00	0.20	0.58	7.28	1.96	81.80	2.08
iFORM-w	0.96	1.00	0.01	0.94	0.83	1.00	0.11	0.50	8.18	1.30	89.02	1.22
iFORM	0.98	1.00	0.01	0.90	0.98	1.00	0.02	0.94	8.05	0.73	92.45	0.70
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	8.00	0.61	93.36	0.18
						$\sigma = 3$						
FS2	0.14	1.00	0.60	0.14	0.03	1.00	0.91	0.02	2.18	6.19	26.54	2.83
iMART1	0.75	1.00	0.07	0.46	0.59	1.00	0.21	0.11	11.95	3.15	70.70	1.95
iMART2	0.73	1.00	0.07	0.71	0.09	1.00	0.45	0.07	8.01	3.62	67.01	1.52
iFORT-w	0.58	1.00	0.18	0.58	0.21	1.00	0.60	0.19	5.21	3.93	59.93	2.27
iFORT	0.58	1.00	0.18	0.57	0.58	1.00	0.32	0.46	6.51	2.91	67.35	2.53
iFORM-w	0.68	1.00	0.15	0.68	0.29	1.00	0.54	0.20	5.77	3.59	62.39	2.32
iFORM	0.92	1.00	0.04	0.86	0.64	1.00	0.19	0.62	7.28	1.82	79.30	1.59
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	8.00	0.91	86.18	0.36
						$\sigma = 4$						
FS2	0.03	1.00	0.76	0.03	0.00	1.00	0.97	0.00	1.24	6.91	11.72	2.24
iMART1	0.67	1.00	0.10	0.41	0.41	1.00	0.31	0.01	12.01	4.00	56.82	1.98
iMART2	0.67	1.00	0.09	0.67	0.05	1.00	0.50	0.03	7.94	4.03	57.53	1.53
iFORT-w	0.29	1.00	0.33	0.29	0.02	1.00	0.89	0.02	3.34	5.33	41.31	1.82
iFORT	0.29	1.00	0.33	0.29	0.27	1.00	0.56	0.22	4.97	4.41	48.54	2.40
iFORM-w	0.24	1.00	0.41	0.24	0.02	1.00	0.89	0.02	3.01	5.62	35.73	2.10
iFORM	0.61	1.00	0.18	0.60	0.15	1.00	0.59	0.15	4.99	3.88	55.68	2.08

NIH-PA Author Manuscript

NIH-PA Author Manuscript

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	**Cor0**	**Inc0**	**Ext**	**iCov**	**iCor0**	**iInc0**	**Iext**	**size**	**MSE**	**Rsq**	**sdR**
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	8.00	1.21	77.73	0.54

Table 3

Results of Example 3, $(n, p, p_0, q_0) = (400, 5000, 10, 10)$, AR correlation.

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	**Cor0**	**Inc0**	**Ext**	**iCov**	**iCor0**	**iInc0**	**Iext**	**size**	**MSE**	**Rsq**	**sdR**
						$\sigma = 2$						
iMART1	1.00	1.00	0.00	1.00	0.69	1.00	0.04	0.69	19.66	1.00	97.76	0.05
iMART2	0.99	1.00	0.00	0.99	0.02	1.00	0.34	0.02	16.89	2.29	95.17	0.20
iFORT-w	0.00	1.00	0.33	0.00	0.03	1.00	0.29	0.03	14.13	5.51	90.89	0.35
iFORT	0.00	1.00	0.33	0.00	0.00	1.00	0.57	0.00	14.80	6.45	86.86	0.55
iFORM-w	1.00	1.00	0.00	1.00	0.93	1.00	0.01	0.62	20.28	0.88	97.91	0.03
iFORM	1.00	1.00	0.00	1.00	0.98	1.00	0.00	0.37	20.74	0.82	97.93	0.03
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.79	97.94	0.03
						$\sigma = 3$						
iMART1	1.00	1.00	0.00	1.00	0.31	1.00	0.10	0.25	19.40	1.49	95.30	0.08
iMART2	1.00	1.00	0.00	1.00	0.02	1.00	0.34	0.02	16.90	2.37	93.00	0.21
iFORT-w	0.00	1.00	0.36	0.00	0.00	1.00	0.37	0.00	13.14	5.97	87.35	0.42
iFORT	0.00	1.00	0.36	0.00	0.00	1.00	0.60	0.00	13.98	6.78	83.70	0.59
iFORM-w	1.00	1.00	0.00	1.00	0.22	1.00	0.14	0.11	19.12	1.70	95.34	0.07
iFORM	1.00	1.00	0.00	1.00	0.37	1.00	0.10	0.15	19.90	1.52	95.50	0.06
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.00	95.79	0.05
						$\sigma = 4$						
iMART1	1.00	1.00	0.00	0.98	0.01	1.00	0.14	0.08	19.66	1.89	92.26	0.14
iMART2	0.97	1.00	0.00	0.97	0.01	1.00	0.36	0.01	16.84	2.56	89.96	0.23
iFORT-w	0.00	1.00	0.40	0.00	0.00	1.00	0.48	0.00	11.50	6.68	82.31	0.41
iFORT	0.00	1.00	0.40	0.00	0.00	1.00	0.66	0.00	12.72	7.38	79.12	0.59
iFORM-w	0.85	1.00	0.02	0.85	0.00	1.00	0.27	0.00	17.57	2.53	91.47	0.17
iFORM	0.94	1.00	0.01	0.94	0.02	1.00	0.22	0.01	18.81	2.24	91.97	0.13
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.24	92.91	0.09

Table 4

Results of Example 4 with ultra-high dimensional data, $(n, p, p_0, q_0) = (400, 10000, 10, 10)$.

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	Cor0	Inc0	Ext	iCov	iCor0	iInc0	Iext	size	MSE	Rsq	sdR
						$\sigma = 2$						
iMART1	0.98	1.00	0.00	0.98	0.54	1.00	0.06	0.54	19.44	1.13	97.63	0.88
iMART2	0.97	1.00	0.00	0.97	0.01	1.00	0.34	0.01	16.84	2.32	94.95	0.21
iFORT-w	0.00	1.00	0.35	0.00	0.00	1.00	0.29	0.00	13.98	5.67	90.50	0.34
iFORT	0.00	1.00	0.35	0.00	0.00	1.00	0.60	0.00	14.55	6.67	86.35	0.54
iFORM-w	1.00	1.00	0.00	0.99	0.95	1.00	0.01	0.68	20.29	0.85	97.91	0.04
iFORM	1.00	1.00	0.00	0.97	0.99	1.00	0.00	0.47	20.66	0.82	97.92	0.03
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.79	97.94	0.03
						$\sigma = 3$						
iMART1	0.98	1.00	1.00	0.98	0.25	1.00	0.11	0.22	19.25	1.55	95.23	0.09
iMART2	0.97	1.00	0.00	0.97	0.02	1.00	0.36	0.02	16.86	2.46	92.78	0.22
iFORT-w	0.00	1.00	0.38	0.00	0.00	1.00	0.40	0.00	12.63	6.22	86.65	0.41
iFORT	0.00	1.00	0.38	0.00	0.00	1.00	0.65	0.00	13.57	7.09	82.92	0.59
iFORM-w	1.00	1.00	0.00	0.99	0.16	1.00	0.16	0.13	18.78	1.77	95.20	0.08
iFORM	1.00	1.00	0.00	0.98	0.35	1.00	0.11	0.18	19.63	1.58	95.39	0.07
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.01	95.78	0.06
						$\sigma = 4$						
iMART1	0.97	1.00	1.00	0.97	0.12	1.00	0.16	0.07	19.66	2.01	92.05	0.13
iMART2	0.97	1.00	0.00	0.97	0.02	1.00	0.36	0.02	16.91	2.59	89.85	0.25
iFORT-w	0.00	1.00	0.42	0.00	0.00	1.00	0.52	0.00	11.03	6.91	81.24	0.48
iFORT	0.00	1.00	0.42	0.00	0.00	1.00	0.68	0.00	12.56	7.50	78.78	0.63
iFORM-w	0.83	1.00	0.02	0.83	0.00	1.00	0.29	0.00	17.44	2.63	91.35	0.14
iFORM	0.97	1.00	0.00	0.97	0.01	1.00	0.23	0.01	18.43	2.26	91.90	0.13
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.25	92.91	0.09

Table 5

Results of Example 5, $(n, p, p_0, q_0) = (400, 10000, 10, 10)$, FS2010 correlation.

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	Cor0	Inc0	Ext	iCov	iCor0	iInc0	Iext	size	MSE	Rsq	sdR
						$\sigma = 2$						
iMART1	0.30	1.00	0.09	0.29	0.03	1.00	0.18	0.24	19.56	2.33	88.90	0.54
iMART2	0.30	1.00	0.09	0.30	0.00	1.00	0.38	0.00	16.69	2.95	85.97	0.55
iFORT-w	0.68	1.00	0.07	0.64	0.75	1.00	0.09	0.73	18.87	1.70	91.59	0.59
iFORT	0.68	1.00	0.07	0.64	0.68	1.00	0.11	0.30	19.48	1.78	91.17	0.67
iFORM-w	1.00	1.00	0.00	0.89	1.00	1.00	0.00	0.90	20.39	0.72	95.12	0.05
iFORM	0.98	1.00	0.01	0.90	0.98	1.00	0.02	0.98	19.98	0.86	94.46	0.47
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.70	95.13	0.05
						$\sigma = 3$						
iMART1	0.28	1.00	0.10	0.21	0.03	1.00	0.20	0.10	20.78	2.66	82.91	0.76
iMART2	0.28	1.00	0.10	0.28	0.00	1.00	0.39	0.00	17.40	3.17	79.98	0.69
iFORT-w	0.60	1.00	0.08	0.60	0.21	1.00	0.24	0.20	17.20	2.40	84.39	0.62
iFORT	0.60	1.00	0.08	0.59	0.60	1.00	0.13	0.29	19.08	2.11	85.32	0.68
iFORM-w	0.96	1.00	0.01	0.93	0.32	1.00	0.15	0.31	18.89	1.60	87.99	0.33
iFORM	0.96	1.00	0.01	0.89	0.64	1.00	0.07	0.63	19.57	1.29	88.67	0.51
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.87	89.98	0.10
						$\sigma = 4$						
iMART1	0.22	1.00	0.13	0.11	0.01	1.00	0.27	0.00	21.06	3.24	74.74	0.76
iMART2	0.22	1.00	0.13	0.22	0.00	1.00	0.43	0.00	17.77	3.61	71.77	0.75
iFORT-w	0.34	1.00	0.13	0.34	0.00	1.00	0.48	0.00	14.35	3.49	73.68	0.66
iFORT	0.34	1.00	0.13	0.32	0.33	1.00	0.21	0.13	17.77	2.85	76.51	0.68
iFORM-w	0.86	1.00	0.04	0.86	0.00	1.00	0.44	0.00	15.83	2.80	77.78	0.53
iFORM	0.91	1.00	0.02	0.91	0.02	1.00	0.34	0.02	16.85	2.41	79.46	0.47
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.07	83.61	0.15

Table 6

Results of Example 6 for the weak heredity case, $(n, p, p_0, q_0) = (400, 5000, 10, 10)$.

	Linear Term Selection				Interaction Selection				Size and Prediction			
	Cov	**Cor0**	**Inc0**	**Ext**	**iCov**	**iCor0**	**iInc0**	**Iext**	**size**	**MSE**	**Rsq**	**sdR**
						$\sigma = 2$						
iMART1	1.00	1.00	0.00	0.92	0.00	1.00	0.56	0.00	16.04	3.68	90.86	0.13
iMART2	1.00	1.00	0.00	0.93	0.00	1.00	0.67	0.00	13.79	3.82	89.99	0.17
iFORT-w	0.04	1.00	0.20	0.04	0.06	1.00	0.24	0.06	15.83	3.95	93.62	0.28
iFORT	0.04	1.00	0.20	0.04	0.00	1.00	0.65	0.00	13.53	5.36	87.11	0.33
iFORM-w	1.00	1.00	0.00	1.00	1.00	1.00	0.00	0.91	20.09	0.76	97.80	0.03
iFORM	0.96	1.00	0.00	0.96	0.00	1.00	0.60	0.00	14.61	3.71	90.73	0.12
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.75	97.81	0.03
						$\sigma = 3$						
iMART1	0.99	1.00	0.00	0.87	0.00	1.00	0.57	0.00	16.58	3.84	88.33	0.17
iMART2	1.00	1.00	0.00	0.85	0.00	1.00	0.67	0.00	13.93	3.87	87.74	0.19
iFORT-w	0.00	1.00	0.24	0.00	0.00	1.00	0.34	0.00	14.51	4.49	89.96	0.33
iFORT	0.00	1.00	0.24	0.00	0.00	1.00	0.68	0.00	12.94	5.70	84.02	0.36
iFORM-w	1.00	1.00	0.00	1.00	0.21	1.00	0.15	0.19	18.66	1.60	94.85	0.07
iFORM	0.79	1.00	0.03	0.79	0.00	1.00	0.63	0.00	14.12	3.96	88.00	0.19
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.97	95.42	0.06
						$\sigma = 4$						
iMART1	0.96	1.00	0.00	0.67	0.00	1.00	0.59	0.00	17.42	4.03	85.09	0.20
iMART2	1.00	1.00	0.00	0.67	0.00	1.00	0.67	0.00	14.30	3.96	84.62	0.22
iFORT-w	0.00	1.00	0.29	0.00	0.00	1.00	0.48	0.00	12.60	5.37	84.16	0.44
iFORT	0.00	1.00	0.29	0.00	0.00	1.00	0.72	0.00	11.91	6.30	79.57	0.43
iFORM-w	0.86	1.00	0.02	0.86	0.00	1.00	0.30	0.00	17.06	2.38	90.77	0.14
iFORM	0.48	1.00	0.08	0.48	0.00	1.00	0.68	0.00	13.05	4.47	83.94	0.29
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.20	92.26	0.09

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 7

Average computation time (in seconds) for $\sigma = 2$.

Example	n	p	(p_0, q_0)	FS2	iFORT	iFORT-w	iFORM	iFORM-w
1	100	500	(4, 4)	16.40	0.04	0.36	0.09	0.51
2	100	500	(4, 4)	16.29	0.04	0.34	0.08	0.50
3	400	5000	(10, 10)	-	11.39	80.66	16.06	126.21
4	400	10000	(10, 10)	-	22.13	144.29	29.17	209.65

Table 8

Prediction performance: the average out-of-sample R^2 for iFOR methods.

Dataset	iFORT	iFORT-w	iFORM	iFORM-w
Inbred Mouse data	60.73 (1.15)	58.46 (1.37)	60.22 (1.15)	60.31 (1.28)
Supermarket data	88.91 (0.17)	88.42 (0.19)	88.66 (0.18)	86.61 (0.22)