Interaction Selection for High Dimensional Data

Hao Helen Zhang

University of Arizona

Workshop on Distributed and Parallel Data Analysis (DPDA)

Motivations for Modeling with Interactions

Given iid data $\{Y_i; X_{i1}, \cdots, X_{ip}\}_{i=1}^n$, a standard linear regression model assumes

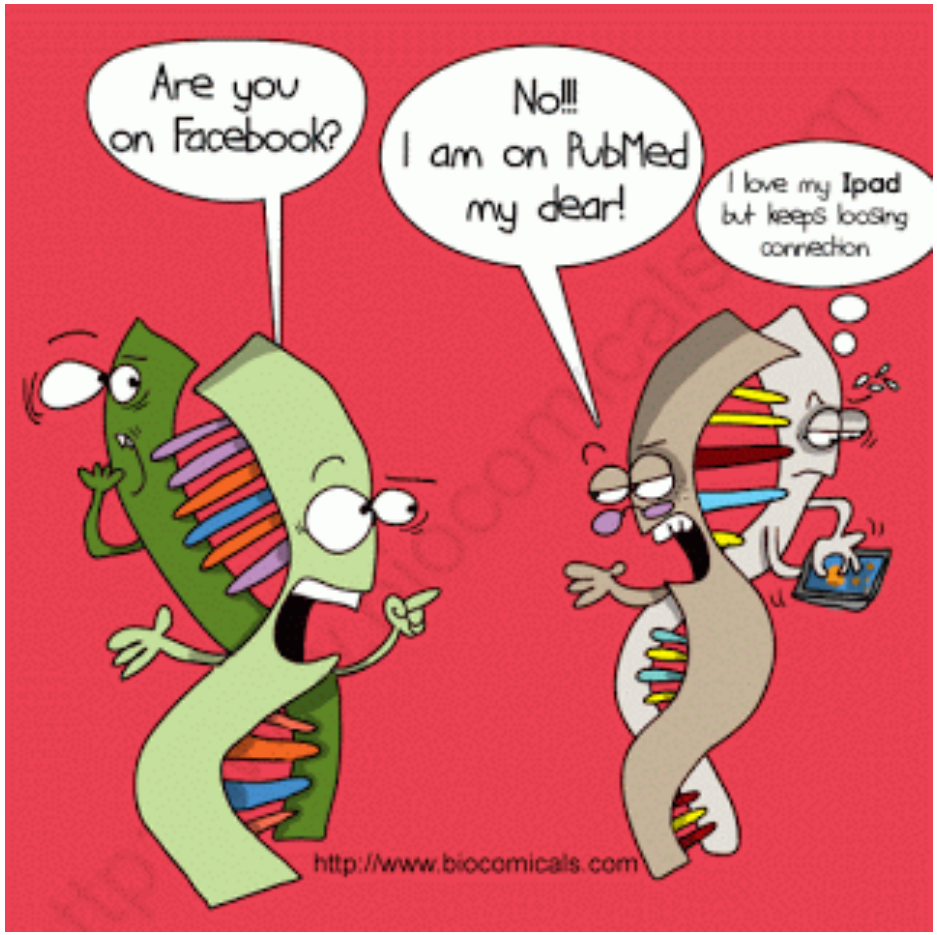
$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, ..., n.$$

There is an increasing need of considering interactions.

- In GWAS, gene-gene (GxG), gene-environmental (GxE) interactions (Manolio et al. 2007; Kooperberg et al. 2008; Cordell 2009).
- In bioassay and epidemiology, the effects of combinations of various behaviors and exposures on disease rates.
- Incidence of lung cancer is accelerated by the combination of smoking and exposure to airborne industrial toxins in a non-linear fashion (Hertz-Picciotto 1992).
- In sociology, interactions between politics and economic growth.

Genes Do Talk to Each Other!

are-you-on-facebook.gif (GIF Image, 320 × 320 pixels) http://1.bp.blogspot.com/-kDAFac



Linear Models with Interactions and Challenges

Consider two-way interactions:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \cdots + \gamma pp X_p^2 + \epsilon.$$

- The total number of parameters is $q = 1 + p + p(p+1)/2$.
- If $p = 10,000$, the number of parameters $\sim O(10^8)$. It is hard to store the entire design matrix.

Computational Challenges.

- For k-way interactions, there are $\binom{p}{k}$ estimaters involved.
- For ultra-high $p \sim e^{n^\xi}$, curse of dimensionality is serious.

Theoretical Challenges

- For a random design, interactions have heavier tails than main effects;
- The correlation structure among interactions is more complex than among main effects.

Hierarchical Model Structures

There is a natural hierarchical structure among predictors,

- X_j and X_k are "parents" of X_jX_k, and X_jX_k is the "child" of X_j and X_k.

Nelder (1977,1994), McCullagh & Nelder (1989), McCullagh (2002) proposed the marginality principle

- ".... X_1 and X_2 must be fitted before X_1X_2".

Strong heredity condition: an interaction effect is selected ONLY IF both of its parents are already selected.

$$\gamma_{jk} \neq 0 \quad \text{only if} \quad \beta_j \neq 0 \text{ and } \beta_k \neq 0.$$

Weak heredity condition: an interaction effect is selected ONLY IF at least one of its parents are already selected.

$$\gamma_{jk} \neq 0 \quad \text{only if} \quad \beta_j \neq 0 \text{ or } \beta_k \neq 0.$$

One-stage Joint Selection via Shrinkage Methods

- Naive methods: ignore the hierarchical structure;
- Hierarchy-enforcement: use asymmetric penalty or linear inequalities to maintain the model hierarchy (Yuan, Joseph & Zou 2009; Zhao, Rocha, & Yu 2009; Choi, Li & Zhu 2010; Bien et al. 2013)

$$\min_{\boldsymbol{\beta}} \quad \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{j,k} \gamma_{jk} x_{ij} x_{ik})^2 + \lambda J(|\beta|),$$

subject to strong hierchichy constraint

Advantages: theoretical justifications, effective for moderate p.
Limitations: require special optimization programming, not feasible for large p.

The largest p used in numerical experiments was $p = 250$.

Scalable Interaction Selection for Ultra-high p

Propose new methods for interaction selection:

- based on forward selection (iFORT, iFORM, iFORM-w)
- based on penalized framework (two-stage LASSO, RAMP)

Advantages: parallel computation, theoretical guarantee, obey hierarchy naturally

Propose new methods for screening interactions (not covered today)

- new correlation measures for interaction effects with the response, after taking into account main effects

Advantages: parallel computation, not require hierarchy constraints

New Definition of "Important Main Effects"

How to decide whether X_j is important for the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \cdots + \gamma pp X_p^2 + \epsilon.$$

- The traditional definition $\beta_j \neq 0$ or $\text{sign}(\beta_j) \neq 0$ is no longer proper, as it violates the invariance principle.
- We need a a new definition.

Definition

We say X_j is important if and only if $\beta_j^2 + \sum_{k=1}^{p} \gamma_{jk}^2 > 0$, and $X_j X_k$ is important if $\gamma_{jk} \neq 0$.

Forward Selection (FS) Algorithm

Let $\widetilde{\mathcal{M}} = \{1, 2, ..., p\}$ and $\widetilde{\mathcal{I}} = \{(k, \ell) : 1 \le k \le \ell \le p\}$. Let

- $\mathcal{S}_k$ denote the index of selected variables after step k.
- $\text{RSS}_{\mathcal{M}}$ denote the residual sum of squares of model $\mathcal{M}$.

Forward Selection (FS) Algorithm:

- Set $k = 0$ and $\mathcal{S}_0 = \emptyset$.
- Let $k = k + 1$. If $k > n$, stop. Otherwise, for each $j \in \widetilde{\mathcal{M}} \backslash \mathcal{S}_{k-1}$, construct a candidate model $\mathcal{M}_{j,k-1} = \mathcal{S}_{k-1} \cup \{j\}$ and compute $\text{RSS}_{\mathcal{M}_{j,k-1}}$. Add the new variable

$$a_k = \arg \min_{j \in \widetilde{\mathcal{M}} \backslash \mathcal{S}_{k-1}} \text{RSS}_{\mathcal{M}_{j,k-1}}$$

and update $\mathcal{S}_k = \mathcal{S}_{k-1} \cup \{a_k\}$. Repeat this step until stop.

Two-stage Approach via Forward Selection (iFORT)

Let $\mathcal{C}$ be a candidate set which contains the effects to be considered for selection.

iFORT Algorithm

Stage 1: Implement FS on $\mathcal{C} = \widetilde{\mathcal{M}}$. Denote the solution path by $\{\mathcal{S}_t^{(1)}, t = 1, 2, \cdots\}$. The set of selected main effects is $\widehat{\mathcal{M}} = \{j_1, \ldots, j_{t_1}\}$.

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

Model Tuning

To select the optimal model from the FS path, we consider BIC.

- Standard BIC

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

- BIC designed for high dimensional data (Chen & Chen, 2008)

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

where d^* is the number of predictors in the full model.

Potential Limitations.

Potential drawbacks of two-stage methods:

- Interactions have to wait until after the main effects have been selected;
- Use stopping rule/tune regularization parameter twice.

Alternatively, we can select the linear and order-2 terms altogether under the marginality principle.

Interaction Selection under Marginality Condition

Main ideas: at each step t, we expand the candidate set $\mathcal{C}_t$ by including interactions between all the main effects in the current model $\mathcal{M}_t$.

$$\mathcal{C}_t = \mathcal{M}_t \cup \{(k, \ell) : k, \ell \in \mathcal{M}_t\}.$$

Advantages:

- an interaction effect is activated immediately after its parents enter the model
- it obeys the hierarchical structure

iFOR under Marginality (iFORM)

iFORM Algorithm

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

Step 2: (Selection) At step t, given $\mathcal{S}_{t-1}$, $\mathcal{C}_{t-1}$ and $\mathcal{M}_{t-1}$, use FS to add one 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 $\mathcal{S}_t$. If the new term is a main effect, we update $\mathcal{C}_t$ and $\mathcal{M}_t$ under the strong heredity. 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 $\{\mathcal{S}_t : t = 1, 2, ...D\}$.

Here $\mathcal{S}_t$, $\mathcal{M}_t$ and $\mathcal{C}_t$ denote the index set of the selected model, the index set of selected main effects, and the candidate set, respectively.

iFORM Under Weak Heredity

iFORM-w Algorithm

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

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 $\mathcal{C}_{t-1} \backslash \mathcal{S}_{t-1}$ into the model. We add the selected one into $\mathcal{S}_{t-1}$ to get $\mathcal{S}_t$. We update $\mathcal{C}_t$ and $\mathcal{M}_t$ if the newly selected predictor is a main effect, under the weak heredity. 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 $\{\mathcal{S}_t : t = 1, 2, ...D\}$.

Two-stage LASSO for Interaction Selection

Stage 1: Solve the standard LASSO

$$\widehat{\boldsymbol{\beta}}_{\text{main}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Define $\widehat{\mathcal{M}} = \text{supp}(\widehat{\boldsymbol{\beta}}_{\text{main}})$.

Stage 2: Minimize

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j \in \widehat{\mathcal{M}}} \beta_j X_j - \sum_{j,k \in \widehat{\mathcal{M}}} \gamma_{jk} X_j X_k \right)^2 + \lambda \sum_{j,k \in \widehat{\mathcal{M}}} |\gamma_{jk}|.$$

Note: At Stage 2, NO penalty is imposed on main effects (to preserve model hierarchy).

Hierarchy-Preserving Solution Path for LASSO

An efficient algorithm to compute a hierarchy-preserving regularization solution path for LASSO.

- implements coordinate descent algorithm (CDA) under the marginality principle.
- Given a tuning parameter λ, the algorithm computes the ℓ_1 regression coefficients of main effects and interactions subject to the heredity condition.

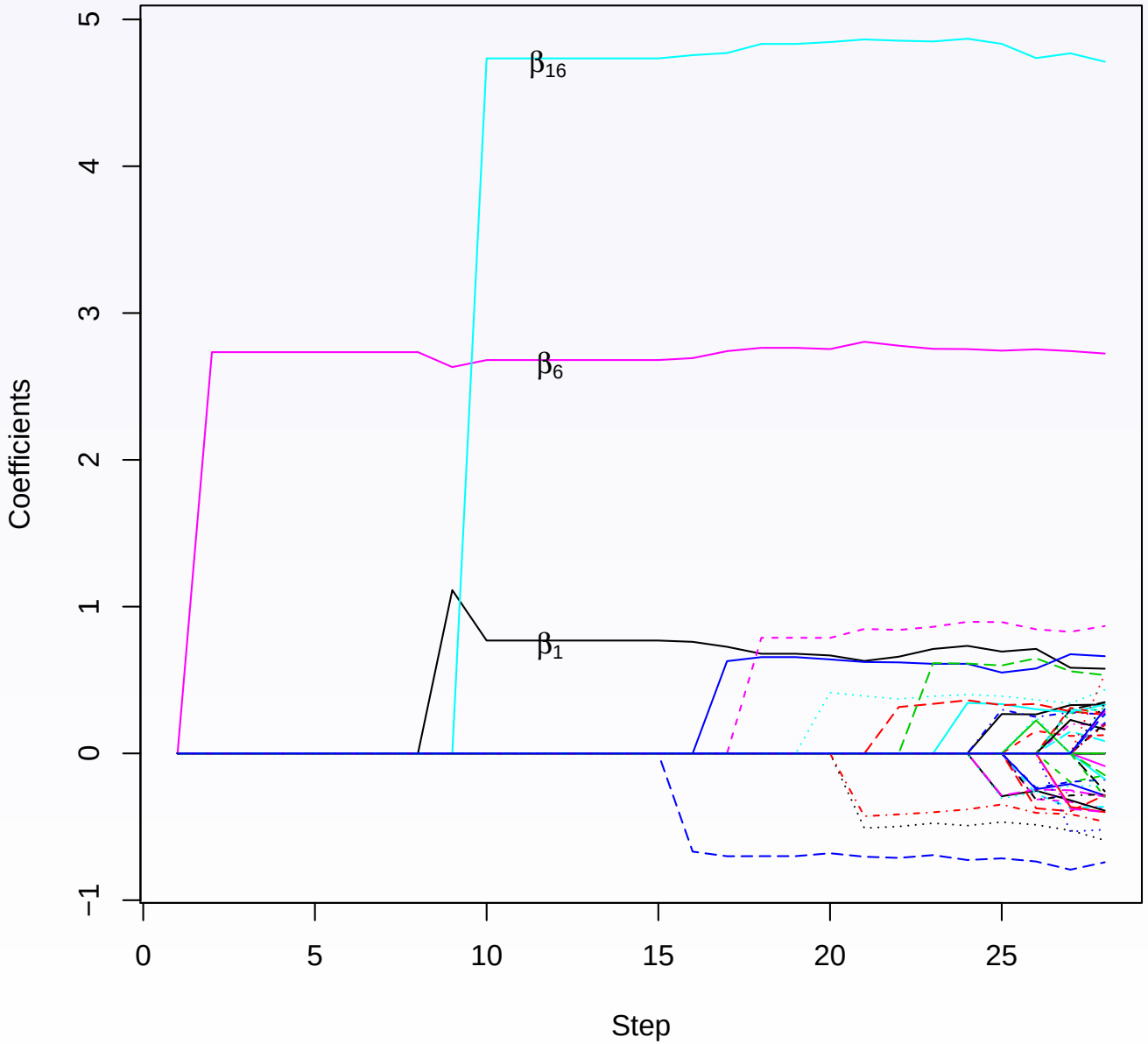
At step $k-1$, denote the current active main effect set as $\mathcal{M}_{k-1}$ and the interaction effect set as $\mathcal{I}_{k-1}$.
Define $\mathcal{H}_{k-1}$ as the parent set of $\mathcal{I}_{k-1}$. Set $\mathcal{H}_{k-1}^c = \widetilde{\mathcal{M}} - \mathcal{H}_{k-1}$.

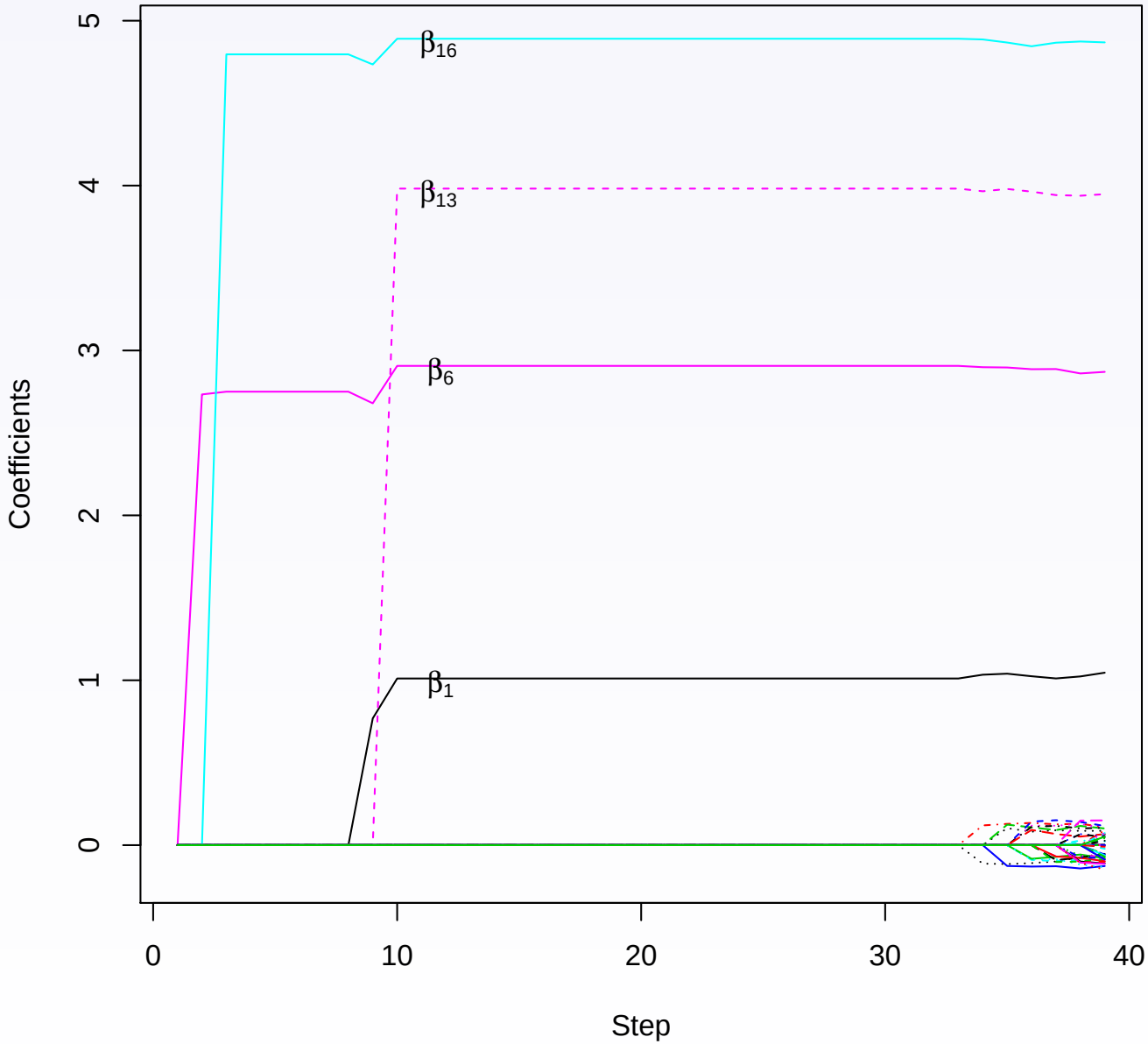
Regularization under Marginality Principle (RAMP)

- (Initialization): Set $\lambda_{\max} = n^{-1} \max |\mathbf{X}^\top \mathbf{y}|$ and $\lambda_{\min} = \zeta \lambda_{\max}$ with some small $\zeta > 0$. Generate an exponentially decaying sequence $\lambda_{\max} = \lambda_1 > \lambda_2 > \cdots > \lambda_K = \lambda_{\min}$. Initialize the main effect set $\mathcal{M}_0 = \emptyset$ and the interaction effect set $\mathcal{I}_0 = \emptyset$.
- (Path-building): Repeat the following for $k = 1, \cdots, K$. Given $\mathcal{M}_{k-1}, \mathcal{I}_{k-1}, \mathcal{H}_{k-1}$, add interactions among main effects in $\mathcal{M}_{k-1}$ to the current model, solve $(\beta_0, \boldsymbol{\beta}_{\mathcal{M}}^\top, \gamma_{\mathcal{M}_{k-1}^{\circ 2}}^\top)^\top$ by minimizing

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}_{\widetilde{\mathcal{M}}} - (\mathbf{x}_i^\top)_{\mathcal{M}_{k-1}}^{\circ 2} \gamma_{\mathcal{M}_{k-1}^{\circ 2}} \right)^2 + \lambda_k (\|\boldsymbol{\beta}_{\mathcal{H}_{k-1}^c}\|_1 + \|\gamma_{\mathcal{M}_{k-1}^{\circ 2}}\|_1).$$

  Record $\mathcal{M}_k$, $\mathcal{I}_k$ and $\mathcal{H}_k$. Add the corresponding main effects from $\mathcal{I}_k$ into $\mathcal{M}_k$ to enforce the heredity constraint, and calculate the OLS based on the current model.


Coefficients
Step
β16
β6
β1
−1
0
1
2
3
4
5
0
5
10
15
20
25


Coefficients
Step
0
1
2
3
4
5
0
10
20
30
40
β16
β13
β6
β1

Scalability of RAMP

RAMP is scalable for large p.

- Unlike existing methods, RAMP avoids storing $O(p^2) \times n$ design matrix
- not involve complex constraints and penalties (but obey model hierarchy)
- The R package RAMP runs well on a desktop.

Example: For a data set with $n = 400$ and $p = 10^4$, it takes less than 30 seconds to obtain the whole solution path (CPU 3.4 GHz Intel Core i7, 32GB memory).

Main Theoretical Results

For two-stage FS method,

- Under some regularity conditions and the strong heredity condition, iFORT is screening consistent.
- Under some regularity conditions and the strong heredity condition, iFORM is screening conistent.

For two-stage LASSO method:

- Under standard conditions for LASSO asymptotics (irrepresentable condition, eigenvalue condition) and the random Gaussian design, two-stage LASSO is sign consistent.

For binary or count data,

- Two-stage LASSO can be extended to GLM model settings.

Regularity Conditions for iFORT

The following regularity conditions are needed.

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

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

C3. Signal strength: We assume that $||\boldsymbol{\beta}|| \leq C_{\boldsymbol{\beta}}$ for some constant $C_{\boldsymbol{\beta}}$ and $\beta_{\min} \geq \nu_\beta n^{-\xi_{\min}}$, where ν_β, $\xi_{\min}$ are positive constants, $\beta_{\min} = \min_{\kappa \in \mathcal{T}} |\beta_\kappa|$.

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

Screening Consistency

Define $K = 2\tau_{\max}\nu C_{\boldsymbol{\beta}}^2 \tau_{\min}^{-2} \nu_\beta^{-4}$.

Theorem

Under conditions (C1)-(C4), the first stage of iFORT is screening consistent for the main effects. For $t_1 \geq K\nu n^{2\xi_0 + 4\xi_{\min}}$,

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

where $\mathcal{T}_1$ is the set of important main effects.

Corollary

Under conditions (C1)-(C4) and strong heredity condition, for $t_2 \geq K\nu n^{2\xi_0 + 4\xi_{\min}}$,

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

In other words, the iFORT algorithm is screening consistent.

About Gaussian Assumptions

Gaussian assumptions can be relaxed to weaker conditions:

(C1)'. X_{ij} is sub-Gaussian marginally, and their joint distribution is symmetric with respect to $\mathbf{0}$.

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

Simulation Settings

Example 1: Let $(n, p, p_0, q_0) = (400, 5000, 10, 10)$. Generate $\mathbf{X}_i$ from MVN with $\text{Cov}(X_j, X_k) = 0.5^{|j-k|}$. The true

$$\boldsymbol{\beta}^{(1)} = (3, 3, 3, 3, 3, 2, 2, 2, 2, 2, \mathbf{0}_{4990}).$$

The nonzero interaction set $\mathcal{T}_2$ is

$$\{(1,2), (1,3), (2,3), (2,5), (3,4), (6,8), (6,10), (7,8), (7,9), (9,10)\},$$

and their coefficients are $(2, 2, 2, 2, 2, 1, 1, 1, 1, 1)$. We generate Y from order-2 model, and $\sigma = 2, 3$ or 4.

Example 2: We increase the dimension $p = 10000$ in Example 3.

Run 100 replications for each setting. Report the average results.

Evaluation Model Selection.

Examine the main effect selection:

- Cov: Coverage probability $\frac{1}{100}\sum_{i=1}^{n} I(\mathcal{T}_1 \subset \widehat{\mathcal{T}}_1^{(i)})$
- Cor0: Percentage of correct zeros
  $\sum_{j=1}^{p} I(\widehat{\beta}_j = 0, \beta_j = 0) / \sum_{j=1}^{p} I(\beta_j = 0)$.
- Inc0: Percentage of incorrect zeros
  $\sum_{j=1}^{p} I(\widehat{\beta}_j = 0, \beta_j \neq 0) / \sum_{j=1}^{p} I(\beta_j \neq 0)$.
- Ext: Exact selection probability $\frac{1}{100}\sum_{i=1}^{n} I(\mathcal{T}_1 = \widehat{\mathcal{T}}_1^{(i)})$.

Examine the interaction selection:

- The counterpart measures iCov, iCor0, iInc0, iExt.

Evaluation Model Fit and Prediction.

For the overall selection, we report

- size:: the selected model size.

For prediction performance,

- MSE: $\sum_{j=1}^{p}(\widehat{\beta}_j - \beta_j)^2 + \sum_{j,k}(\widehat{\gamma}_{jk} - \gamma_{jk})^2)$.
- Rsq: Out-of-sample R^2

$$100\% \times \{1 - \frac{\sum_{i=1}^{n^*}(Y_i^* - x_i^*\hat{\beta}^{(1)} - z_i^*\hat{\beta}^{(2)})^2}{\sum_{i=1}^{n^*}(Y_i^* - \bar{Y}^*)^2}\},$$

  $(\mathbf{X}_i^*, Y_i^*), i = 1, \cdots, n^*$ are test data, independently from the same distribution as the training set, $\bar{Y}^* = \frac{1}{n^*}\sum_{i=1}^{n^*} Y_i^*$.

- sdR: the standard error of Rsq.

Results

Table : Example 1: $(n, p, p_0, q_0) = (400, 5000, 10, 10)$.

	Linear Term Selection				Interaction Selection				Prediction		
	Cov	Cor0	Inc0	Ext	iCov	iCor0	iInc0	Iext	size	MSE	Rsq
						$\sigma = 2$					
iFORT	0.00	1.00	0.33	0.00	0.00	1.00	0.57	0.00	14.80	6.45	86.86
iFORM	1.00	1.00	0.00	1.00	0.98	1.00	0.00	0.37	20.74	0.82	97.93
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.79	97.94
						$\sigma = 3$					
iFORT	0.00	1.00	0.36	0.00	0.00	1.00	0.60	0.00	13.98	6.78	83.70
iFORM	1.00	1.00	0.00	1.00	0.37	1.00	0.10	0.15	19.90	1.52	95.50
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.00	95.79
						$\sigma = 4$					
iFORT	0.00	1.00	0.40	0.00	0.00	1.00	0.66	0.00	12.72	7.38	79.12
iFORM	0.94	1.00	0.01	0.94	0.02	1.00	0.22	0.01	18.81	2.24	91.97
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.24	92.91

Results

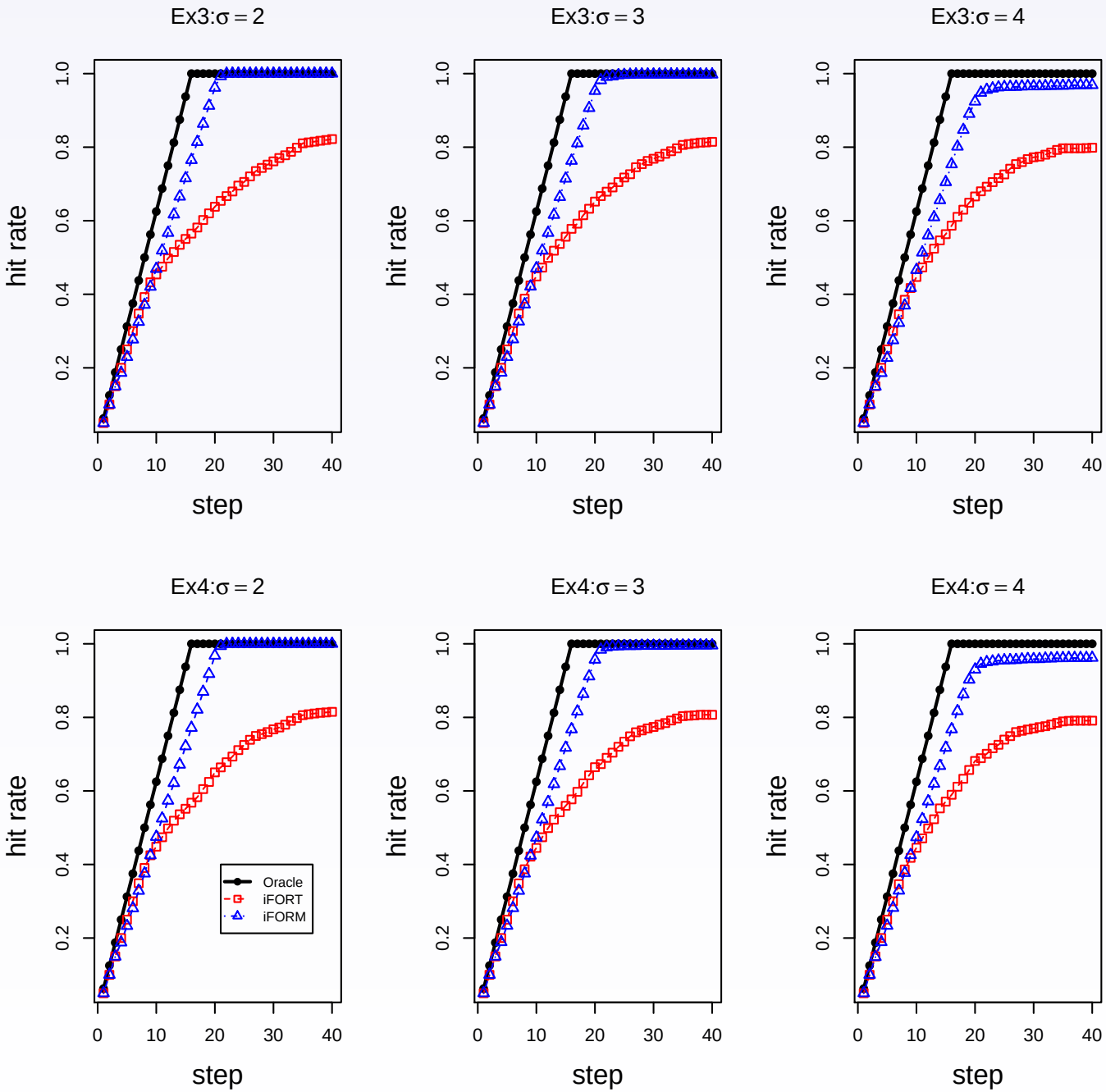
Table : Example 2: $(n, p, p_0, q_0) = (400, 10000, 10, 10)$.

	Linear Term Selection				Interaction Selection				Prediction		
	Cov	Cor0	Inc0	Ext	iCov	iCor0	iInc0	Iext	size	MSE	Rsq
						$\sigma = 2$					
iFORT	0.00	1.00	0.35	0.00	0.00	1.00	0.60	0.00	14.55	6.67	86.35
iFORM	1.00	1.00	0.00	0.97	0.99	1.00	0.00	0.47	20.66	0.82	97.92
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	0.79	97.94
						$\sigma = 3$					
iFORT	0.00	1.00	0.38	0.00	0.00	1.00	0.65	0.00	13.57	7.09	82.92
iFORM	1.00	1.00	0.00	0.98	0.35	1.00	0.11	0.18	19.63	1.58	95.39
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.01	95.78
						$\sigma = 4$					
iFORT	0.00	1.00	0.42	0.00	0.00	1.00	0.68	0.00	12.56	7.50	78.78
iFORM	0.97	1.00	0.00	0.97	0.01	1.00	0.23	0.01	18.43	2.26	91.90
ORACL	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	20.00	1.25	92.91

Path

To better demonstrate the quality of the solution path, we further plot the "hit rate".

1. x-axis is the model size, $1 \leq s \leq S$.
2. y-axis is the so-called "hit rate", representing the percentage of important terms found the first s selected terms.


Ex3:σ = 2
Ex3:σ = 3
Ex3:σ = 4
Ex4:σ = 2
Ex4:σ = 3
Ex4:σ = 4
hit rate
step
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Oracle
iFORT
iFORM

Computation Time

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

Example	p	(p_0, q_0)	FS2	iFORT	iFORM
1	500	(4,4)	16.40	0.04	0.09
2	500	(4,4)	16.29	0.04	0.08
3	5000	(10,10)	-	11.39	16.06
4	10000	(10,10)	-	22.13	29.17

A Binary Example

Consider a logistic regression model with

$$\log \frac{P(Y=1|\mathbf{X})}{P(Y=0|\mathbf{X})} = \beta_1 X_1 + 3X_6 + 3X_{10} + 3X_1X_6 + 3X_6X_{10},$$

with $(n, p, p_0, q_0) = (400, 2000, 3, 2)$, and $\mathbf{X} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$.

- For different signal-to-noise ratios, we vary the coefficient $\beta_1 \in \{1, 2, 3\}$.

Compare RAMP, two-stage LASSO, two-stage SCAD, Oracle.

	β_1	main effects cover	main effects exact	interactions cover	interactions exact	size	RMSE
RAMP	1	0.92	0.78	0.92	0.91	4.98	1.80
	2	1.00	0.93	1.00	1.00	5.08	1.16
	3	1.00	0.92	0.99	0.99	5.13	1.36
2-LASSO	1	0.45	0.41	0.45	0.14	4.05	3.97
	2	1.00	0.93	1.00	0.29	6.58	1.41
	3	1.00	0.80	1.00	0.42	6.31	1.66
2-SCAD	1	0.49	0.43	0.49	0.49	3.58	3.76
	2	1.00	0.81	1.00	0.94	5.28	1.03
	3	1.00	0.74	1.00	0.86	5.52	1.22
ORACLE	1	1.00	1.00	1.00	1.00	5.00	0.84
	2	1.00	1.00	1.00	1.00	5.00	0.78
	3	1.00	1.00	1.00	1.00	5.00	0.83

Weak-Hierarchy Example

Consider the regression model with order-2 interaction.

- Generate $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\Sigma_{jk} = 0.5^{|j-k|}$.
- Let $(n, p, p_0, q_0) = (400, 100, 10, 10)$.
- The index of main effects is $\mathcal{S} = \{1, 2, \cdots, 10\}$ with

$$\boldsymbol{\beta}_{\mathcal{S}} = (3, 3, 3, 3, 3, 2, 2, 2, 2, 2)^{\top}.$$

- The set of important interaction effects is

$$\{(1,2), (1,13), (2,3), (2,15), (3,4), (6,10), (6,18), (7,9), (7,18), (10,19)\}$$

with the corresponding coefficients $(2, 2, 2, 2, 2, 1, 1, 1, 1, 1)$.

- Strong heredity is violated, but weak heredity holds.

We compare with hierNet-s and hierNet-w (Bien et al . 2014).

		main effects		interactions				
	σ	cover	exact	cover	exact	size	RMSE	Time
RAMP	2	1.00	0.71	0.00	0.00	19.45	3.54	37.49
	3	1.00	0.83	0.00	0.00	16.86	3.71	34.74
	4	0.98	0.89	0.00	0.00	15.28	3.87	34.88
RAMP-w	2	1.00	1.00	0.99	0.25	21.33	0.79	47.02
	3	1.00	0.99	0.63	0.12	21.16	1.31	46.51
	4	1.00	0.98	0.16	0.00	20.07	1.98	46.10
hierNet-s	2	1.00	0.00	1.00	0.00	133.45	5.69	3143.30
	3	1.00	0.00	0.96	0.00	119.62	5.33	3232.62
	4	1.00	0.00	0.74	0.00	95.06	5.01	3507.85
hierNet-w	2	1.00	0.00	1.00	0.00	126.83	6.60	295.88
	3	1.00	0.01	0.98	0.00	96.59	6.17	346.83
	4	1.00	0.04	0.75	0.00	65.31	5.73	444.99

Real Data Analysis

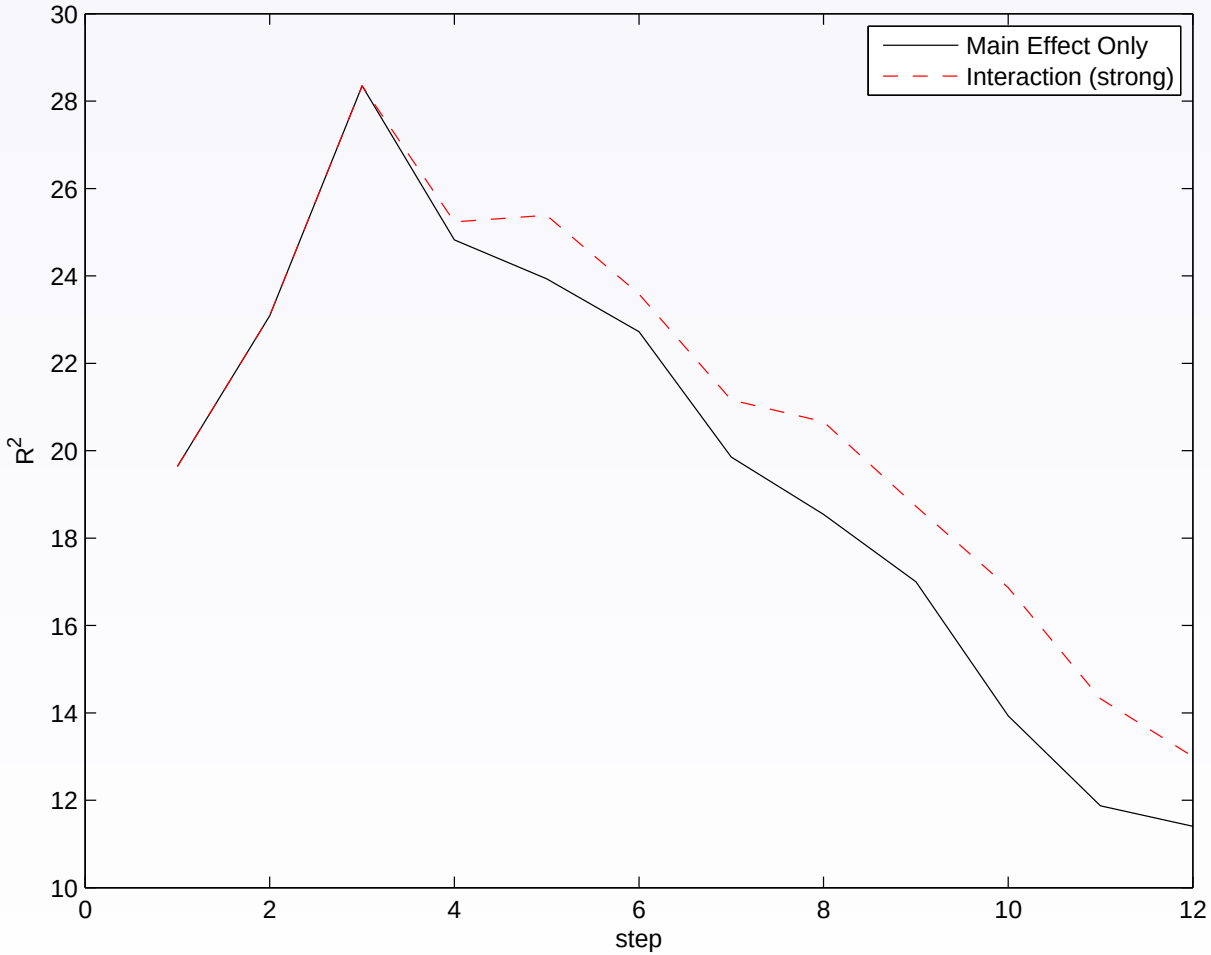
The inbred mouse microarray data (Lan et al. 2006).

- 60 arrays, with 31 from female mice and 29 from male mice, respectively.
- Each array measures the expression values of 22,690 genes.
- Y is a phenotypic variable measured by real-time RT-PCR, stearoyl-CoA desaturase 1 (SCD1).

Use the SIS to pre-select genes with absolute marginal correlation higher than 0.3. This leaves us 1,856 genes.

- iFORT identifies three linear effects probe id 1415742_at, 1434185_at, 1441881_x_at, two interactions 1415742_at $*$ 1441881_x_at, 1434185_at $*$ 1441881_x_at, and one quadratic effect 1434185_at^2.
- iFORM identifies the main effects for the same three genes but not their interactions.

Average Leave-10-out Out-of-sample R^2 with Precreening



References

- Bien, J., Taylor, J. & Tibshirani, R. (2013). A lasso for hierarchical interactions. Annals of Statistics 41, 1111-1141.
- Choi, N. H., Li, W. & Zhu, J. (2010). Variable selection with the strong heredity con- straint and its oracle property. JASA 105, 354-364.
- Hao and Zhang (2014) Interaction screening for ultra-high dimensional data. JASA 109, 1285-1301.
- Hao and Zhang (2016) A note on linear model with interactions. Tentatively accepted.
- Hao, Feng, and Zhang, (2016) Model selection for high dimensional quadratic regression via regularization. Revised and resubmitted.
- Niu, Hao, and Zhang (2016) Screening for interaction effects. Submitted.