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

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

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/-kDAFacINWl0/ToKW_1fnoTI/AA...
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 \( \binom{p}{k} \) estimators 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 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 $X_k$.


- “... $X_1$ 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)

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

subject to strong hierarchy 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 $p = 250$. 
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 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 $\tilde{M} = \{1, 2, ..., p\}$ and $\tilde{I} = \{(k, \ell) : 1 \leq k \leq \ell \leq p\}$. Let

- $S_k$ 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 $S_0 = \emptyset$.
- Let $k = k + 1$. If $k > n$, stop. Otherwise, for each $j \in \tilde{M} \setminus S_{k-1}$, construct a candidate model $\mathcal{M}_{j,k-1} = S_{k-1} \cup \{j\}$ and compute $\text{RSS}_{\mathcal{M}_{j,k-1}}$. Add the new variable

\[
a_k = \arg\min_{j \in \tilde{M} \setminus S_{k-1}} \text{RSS}_{\mathcal{M}_{j,k-1}}
\]

and update $S_k = S_{k-1} \cup \{a_k\}$. 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 $\mathcal{C} = \tilde{\mathcal{M}}$. Denote the solution path by $\{S_{t}^{(1)}, t = 1, 2, \cdots \}$. The set of selected main effects is $\hat{\mathcal{M}} = \{j_1, \ldots, j_{t_1} \}$.

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

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

- Standard BIC

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

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

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

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.
Main ideas: at each step $t$, we expand the candidate set $C_t$ by including interactions between all the main effects in the current model $M_t$.

$$C_t = M_t \cup \{(k, \ell) : k, \ell \in M_t\}.$$ 

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 $S_0 = \emptyset$, $M_0 = \emptyset$ and $C_0 = \tilde{\mathcal{M}}$.

Step 2: (Selection) At step $t$, given $S_{t-1}$, $C_{t-1}$ and $M_{t-1}$, use FS to add one predictor from $C_{t-1} \setminus S_{t-1}$ into the model. We add the selected one into $S_{t-1}$ to get $S_t$. If the new term is a main effect, we update $C_t$ and $M_t$ under the strong heredity. Otherwise, $C_t = C_{t-1}$ and $M_t = M_{t-1}$.

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

Here $S_t$, $M_t$ and $C_t$ 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 $S_0 = \emptyset$, $M_0 = \emptyset$ and $C_0 = \tilde{M}$.

Step 2: (Selection) In the $t$th step with given $S_{t-1}$, $C_{t-1}$ and $M_{t-1}$, forward regression is used to select one more predictor from $C_{t-1} \setminus S_{t-1}$ into the model. We add the selected one into $S_{t-1}$ to get $S_t$. We update $C_t$ and $M_t$ if the newly selected predictor is a main effect, under the weak heredity. Otherwise, $C_t = C_{t-1}$ and $M_t = M_{t-1}$.

Step 3: (Solution path) Iterating Step 2, we get a solution path $\{S_t : t = 1, 2, \ldots D\}$. 
Two-stage LASSO for Interaction Selection

Stage 1: Solve the standard LASSO

\[ \hat{\beta}_{\text{main}} = \arg\min_{\beta} \sum_{i=1}^{n} \left( Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \]

Define \( \hat{M} = \text{supp}(\hat{\beta}_{\text{main}}) \).

Stage 2: Minimize

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

**Note:** At Stage 2, NO penalty is imposed on main effects (to preserve model hierarchy).
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 $\lambda$, the algorithm computes the $\ell_1$ 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 $\mathcal{H}^c_{k-1} = \tilde{\mathcal{M}} - \mathcal{H}_{k-1}$. 
Regularization under Marginality Principle (RAMP)

- **(Initialization):** Set $\lambda_{\text{max}} = n^{-1} \max |X^\top y|$ and $\lambda_{\text{min}} = \zeta \lambda_{\text{max}}$ with some small $\zeta > 0$. Generate an exponentially decaying sequence $\lambda_{\text{max}} = \lambda_1 > \lambda_2 > \cdots > \lambda_K = \lambda_{\text{min}}$. 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 $$(\beta_0, \beta_{\mathcal{M}}^\top, \gamma_{\mathcal{M}_{k-1}^o2}^\top)^\top$$ by minimizing

$$
\sum_{i=1}^{n} \left( y_i - \beta_0 - x_i^\top \beta_{\mathcal{M}} - (x_i^\top)_{\mathcal{M}_{k-1}} \gamma_{\mathcal{M}_{k-1}^o2} \right)^2 + \lambda_k (\| \beta_{\mathcal{H}_{k-1}}^\top \|_1 + \| \gamma_{\mathcal{M}_{k-1}^o2} \|_1).
$$

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.
Figure: LASSO Path for Interactions under Strong Heredity

Step
Coefficients
β_{16}
β_{16}
β_{1}
β_{1}

Step
Coefficients
β_{6}
β_{6}

Hao Helen Zhang Interaction Selection 19/40
Figure: LASSO Path for Interactions under Weak Heredity

Hao Helen Zhang
Interaction Selection
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 consistent.

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
$$2\tau_{\min} < \lambda_{\min}(\Sigma^{(1)}) \leq \lambda_{\max}(\Sigma^{(1)}) < \frac{\tau_{\max}}{2}.$$

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

C4. Dimensionality and sparsity: There exists positive constants $\xi$, $\xi_0$ and $\nu$, 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 $K = 2\tau_{\text{max}} \nu C^2_{\beta} \tau_{\text{min}}^{-2} \nu_{\beta}^{-4}$.

Theorem

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

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

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

Corollary

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

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

In other words, the iFORT algorithm is screening consistent.
Gaussian assumptions can be relaxed to weaker conditions:

(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.
Simulation Settings

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

\[ \beta^{(1)} = (3, 3, 3, 3, 3, 2, 2, 2, 2, 0_{4990}). \]

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

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: $\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)^2 + \sum_{j,k}(\hat{\gamma}_{jk} - \gamma_{jk})^2$.

■ Rsq: Out-of-sample $R^2$

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

$(X_i^*, Y_i^*), i = 1, \cdots, n^*$ 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.
Table: Example 1: \((n, p, p_0, q_0) = (400, 5000, 10, 10)\).

<table>
<thead>
<tr>
<th></th>
<th>Linear Term Selection</th>
<th></th>
<th>Interaction Selection</th>
<th></th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cov</td>
<td>Cor0</td>
<td>Inc0</td>
<td>Ext</td>
<td>iCov</td>
</tr>
<tr>
<td>iFORT</td>
<td>0.00</td>
<td>1.00</td>
<td>0.33</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>iFORM</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>ORACL</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

\(\sigma = 2\)

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>iFORT</td>
<td>0.00</td>
<td>1.00</td>
<td>0.36</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.60</td>
<td>0.00</td>
<td>13.98</td>
<td>6.78</td>
<td>83.70</td>
</tr>
<tr>
<td>iFORM</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.37</td>
<td>1.00</td>
<td>0.10</td>
<td>0.15</td>
<td>19.90</td>
<td>1.52</td>
<td>95.50</td>
</tr>
<tr>
<td>ORACL</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>20.00</td>
<td>1.00</td>
<td>95.79</td>
</tr>
</tbody>
</table>

\(\sigma = 3\)

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>iFORT</td>
<td>0.00</td>
<td>1.00</td>
<td>0.40</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.66</td>
<td>0.00</td>
<td>12.72</td>
<td>7.38</td>
<td>79.12</td>
</tr>
<tr>
<td>iFORM</td>
<td>0.94</td>
<td>1.00</td>
<td>0.01</td>
<td>0.94</td>
<td>0.02</td>
<td>1.00</td>
<td>0.22</td>
<td>0.01</td>
<td>18.81</td>
<td>2.24</td>
<td>91.97</td>
</tr>
<tr>
<td>ORACL</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>20.00</td>
<td>1.24</td>
<td>92.91</td>
</tr>
</tbody>
</table>

\(\sigma = 4\)
## Results

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

<table>
<thead>
<tr>
<th></th>
<th>Linear Term Selection</th>
<th></th>
<th>Interaction Selection</th>
<th></th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cov</td>
<td>Cor0</td>
<td>Inc0</td>
<td>Ext</td>
<td>iCov</td>
</tr>
<tr>
<td>iFORT</td>
<td>0.00</td>
<td>1.00</td>
<td>0.35</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>iFORM</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>ORACL</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iFORT</td>
<td>0.00</td>
<td>1.00</td>
<td>0.38</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>iFORM</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.98</td>
<td>0.35</td>
</tr>
<tr>
<td>ORACL</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iFORT</td>
<td>0.00</td>
<td>1.00</td>
<td>0.42</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>iFORM</td>
<td>0.97</td>
<td>1.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.01</td>
</tr>
<tr>
<td>ORACL</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
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.
Table: Average computation time (in seconds) for $\sigma = 2$.

<table>
<thead>
<tr>
<th>Example</th>
<th>$p$</th>
<th>$(p_0, q_0)$</th>
<th>FS2</th>
<th>iFORT</th>
<th>iFORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>500</td>
<td>(4,4)</td>
<td>16.40</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>(4,4)</td>
<td>16.29</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>5000</td>
<td>(10,10)</td>
<td>-</td>
<td>11.39</td>
<td>16.06</td>
</tr>
<tr>
<td>4</td>
<td>10000</td>
<td>(10,10)</td>
<td>-</td>
<td>22.13</td>
<td>29.17</td>
</tr>
</tbody>
</table>
A Binary Example

Consider a logistic regression model with

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

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

- 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.
<table>
<thead>
<tr>
<th></th>
<th>( \beta_1 )</th>
<th>main effects</th>
<th>interactions</th>
<th>size</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cover</td>
<td>exact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAMP</td>
<td>1</td>
<td>0.92</td>
<td>0.78</td>
<td>4.98</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>0.93</td>
<td>5.08</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>0.92</td>
<td>5.13</td>
<td>1.36</td>
</tr>
<tr>
<td>2-LASSO</td>
<td>1</td>
<td>0.45</td>
<td>0.41</td>
<td>4.05</td>
<td>3.97</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>0.93</td>
<td>6.58</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>0.80</td>
<td>6.31</td>
<td>1.66</td>
</tr>
<tr>
<td>2-SCAD</td>
<td>1</td>
<td>0.49</td>
<td>0.43</td>
<td>3.58</td>
<td>3.76</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>0.81</td>
<td>5.28</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>0.74</td>
<td>5.52</td>
<td>1.22</td>
</tr>
<tr>
<td>ORACLE</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>5.00</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>5.00</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
<td>5.00</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Weak-Hierarchy Example

Consider the regression model with order-2 interaction.

- Generate $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{jk} = 0.5^{|j-k|}$.
- Let $(n, p, p_0, q_0) = (400, 100, 10, 10)$.
- The index of main effects is $S = \{1, 2, \cdots, 10\}$ with $\beta_S = (3, 3, 3, 3, 3, 2, 2, 2, 2, 2)^\top$.
- 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 $(2, 2, 2, 2, 1, 1, 1, 1, 1, 1)$.
- Strong heredity is violated, but weak heredity holds.

We compare with hierNet-s and hierNet-w (Bien et al. 2014).
<table>
<thead>
<tr>
<th></th>
<th>$\sigma$</th>
<th>main effects</th>
<th>interactions</th>
<th>size</th>
<th>RMSE</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cover</td>
<td>exact</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAMP</td>
<td>2</td>
<td>1.00</td>
<td>0.71</td>
<td>0.00</td>
<td>0.00</td>
<td>19.45</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>0.83</td>
<td>0.00</td>
<td>0.00</td>
<td>16.86</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.98</td>
<td>0.89</td>
<td>0.00</td>
<td>0.00</td>
<td>15.28</td>
</tr>
<tr>
<td>RAMP-w</td>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.25</td>
<td>21.33</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>0.99</td>
<td>0.63</td>
<td>0.12</td>
<td>21.16</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.00</td>
<td>0.98</td>
<td>0.16</td>
<td>0.00</td>
<td>20.07</td>
</tr>
<tr>
<td>hierNet-s</td>
<td>2</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>133.45</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>0.00</td>
<td>0.96</td>
<td>0.00</td>
<td>119.62</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.00</td>
<td>0.00</td>
<td>0.74</td>
<td>0.00</td>
<td>95.06</td>
</tr>
<tr>
<td>hierNet-w</td>
<td>2</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>126.83</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.00</td>
<td>0.01</td>
<td>0.98</td>
<td>0.00</td>
<td>96.59</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.00</td>
<td>0.04</td>
<td>0.75</td>
<td>0.00</td>
<td>65.31</td>
</tr>
</tbody>
</table>
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 $\times$ 1441881_x_at, 1434185_at $\times$ 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

![Graph showing $R^2$ values for different steps in the process with two lines: one for Main Effect Only and another for Interaction (strong).]
References