Interaction Screeening for high dimensional data

Yongdai Kim

Department of Statistics
Seoul National University

Outline

1. Introduction

Linear model

- Y: Response
- $X = (X_1, \ldots, X_p)'$:: p-dimensional vectors of covariate.
- Assume that
  - $\mathrm{E}(Y) = 0$
  - $\mathrm{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$.
- Linear regression model is

$$Y = \sum_{j=1}^{p} X_j \beta_j^* + \epsilon, \ \epsilon \sim (0, \sigma^2)$$

- Data: $(Y_1, X_1), \ldots, (Y_n, X_n) \overset{i.i.d.}{\sim} (Y, X)$.

Variable selection

- Let $\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$ be the set of signal covariates.
- The aim of variable selection is to estimate $\mathcal{A}^*$.
- Two regimes of variable selection methods
  - Search: Stepwise selection, Forward selection, SIS (Sure Independent Screening),...
  - Sparse regularization: Lasso, SCAD, bridge,...

Selection consistency versus Screening property

- Let $\hat{\mathcal{A}}_n$ be an estimator of $\mathcal{A}^*$ with data of size n.
- We say that $\hat{\mathcal{A}}_n$ is **selection consistent** if

$$\Pr\left(\hat{\mathcal{A}}_n = \mathcal{A}^*\right) \to 1$$

as $n \to \infty$.

- We say that $\hat{\mathcal{A}}_n$ has **screening property** if

$$\Pr\left(\mathcal{A}^* \subset \hat{\mathcal{A}}_n\right) \to 1$$

as $n \to \infty$.

Selection consistency versus Screening property

- Selection consistent algorithms: nonconvex penalized estimators such as SCAD, MCP, bridge ....
- Algorithms having screening property: Forward selection, Lasso, SIS,...

Practical use of screening property

- Apply an algorithm with screening property to reduce the number of candidate covariates from p to q such that $q << p$.
- Apply a selection consistent algorithm to data only with q covariates selected by the algorithm with screening property.
- An algorithm with screening property is useful only when computation is simple and $\hat{A}_n$ is small:
  - $|\hat{A}_n| = O(n^a)$ for $a > 0$ when $p = O(e^{bn})$ for $b > 0$

Algorithm with screening property: SIS

- Let $w_j = \mathrm{cov}_n(Y, X_j)$, where $\mathrm{cov}_n(Y, X_j)$ is the empirical covariance between Y and X_j with the sample size n.
- Let $\hat{\mathcal{A}}_n = \{j : |w_j| > \tau_n\}$ for some threshold $\tau_n > 0$.
- Under regularity conditions, $\hat{A}_n$ has a screening property provided

$$|\mathrm{cov}_\infty(Y, X_j)| \text{ is large for all } j \in \mathcal{A}^*$$

and

$$|\mathrm{cov}_\infty(Y, X_j)| \text{ is small for most } j \in \mathcal{A}^{*c}.$$

- Fan and Lv [2008]

Algorithm with screening property: Forward selection

- Initialization: $\hat{\mathcal{A}}_n = \emptyset, \beta = 0$
- Repeat until $\hat{\mathcal{A}}_n = \{1, \ldots, p\}$
  - $R = Y - X'\beta$
  - Choose $j = \text{argmax}_k |\text{cov}_n(R, X_k)|$
  - $\hat{\mathcal{A}}_n = \hat{\mathcal{A}}_n \cup \{j\}$ and

$$\beta = \text{argmin}_{\beta:\beta_j=0, j\in\hat{\mathcal{A}}_n^c} \| Y - X'\beta \|_n^2$$

- Wang [2009]

Algorithm with screening property: LARS (Efron et al. [2004])

- Forward-stagewise algorithm
  - Initialization: $\hat{\mathcal{A}}_n = \emptyset, \beta = 0$ and $\gamma > 0$.
  - Repeat until $\hat{\mathcal{A}}_n = \{1, \ldots, p\}$
    - $R = Y - X' \beta$
    - Choose $j = \text{argmax}_k |\text{cov}_n(R, X_k)|$
    - $\hat{\mathcal{A}}_n = \hat{\mathcal{A}}_n \cup \{j\}$ and $\beta_j = \beta_j + \gamma \text{sign}\{\text{cov}_n(R, X_j)\}$
- If $\gamma \to 0$ and adding a deletion step, the forward-stagewise algorithm essentially solves the LASSO problem.

Two way interaction model

- Model

$$Y = \beta_0^* + \sum_{j=1}^{p} \beta_j^* X_j + \sum_{j<k} \beta_{jk}^* X_j X_k + \epsilon$$

  where $\epsilon \sim (0, \sigma^2)$.
- Our goal is to select a subset of the $p + p(p-1)/2$ many main and interaction effect terms.
- When p is large, there are too many terms in the model and brute-forth search may not be computationally feasable.
- **We need a way to reduce computational burden!**

Aim of the talk

- The aim of this talk is **to propose methods of interaction screening whose computational complexity is not much larger than those for linear models.**
- Two assumptions for interaction screening:
  - **heredity**: interactions are significant only when the corresponding main effects are significant (well known)
  - **Gaussianity**: the covariates are Gaussian (new and main them of this talk)

2. Interaction screening under heredity assumption

Heredity of regression coefficients I

- In the words of Cox [1984]:
  "Large component main effects are more likely to lead to appreciable interactions than small components. Also, the interactions corresponding to larger main effects may be in some sense of more practical importance".
- It may be useful to focus our search on interactions that have large main effects.
- Such restrictions are known under various names, including "heredity", "marginality",and being "hierarchically well -formulated".

Heredity of regression coefficients II

- Strong and Weak hierarchy (e,g, Bien et al. [2013])
- Strong heredity

$$\beta_{jk} \neq 0 \text{ only if } \beta_j \neq 0 \textit{ and } \beta_k \neq 0$$

- Weak heredity

$$\beta_{jk} \neq 0 \text{ only if } \beta_j \neq 0 \textit{ or } \beta_k \neq 0$$

SIS for interaction with heredity

- First, select main effects by us of the standard SIS.
- Second, select interactions among those whose corresponding main effects are selected.
- Computational cost is $O(qp)$, where q is the number of selected main effects, which is much smaller than p^2 when $q << p$.

Forward selection for interaction with heredity I

- Consider strong heridity for simplicity.
- For given $\mathcal{A} \subset \{1, \ldots, p\} \cup \{1, \ldots, p\}^2$, let $\mathcal{A}^{(1)} = \mathcal{A} \cap \{1, \ldots, p\}$ and $\mathcal{A}^{(2)} = \mathcal{A} \cap \{1, \ldots, p\}^{(2)}$.
- For given $\mathcal{A} \subset \{1, \ldots, p\} \cup \{1, \ldots, p\}^2$. let

$$\mathcal{I}_{\mathcal{A}} = \{(j, k) : j \in \mathcal{A}^{(1)}, k \in \mathcal{A}^{(1)}\}.$$

- Let $Z = [X_j X_k]_{(j,k) \in \{1,\ldots,p\}^2}$ and $B = [\beta_{jk}]_{(j,k) \in \{1,\ldots,p\}^2}$.

Forward selection for interaction with heredity II

- Initialization: $\hat{\mathcal{A}}_n = \emptyset, \beta = 0$
- Repeat
  - $R = Y - X'\beta - Z'B$
  - Choose $j = \text{argmax}_{k \in \{1,\ldots,p\} - \hat{\mathcal{A}}_n^{(1)}} |\text{cov}_n(R, X_k)|$
  - Choose $(l, h) = \text{argmax}_{(j,k) \in \mathcal{I}_{\hat{\mathcal{A}}_n} - \hat{\mathcal{A}}_n^{(2)}} |\text{cov}_n(R, X_j X_k)|$
  - Let $r = j$ if $|\text{cov}_n(R, X_j)| \geq |\text{cov}_n(R, X_l X_h)|$ or let $r = (l, h)$ otherwise.
  - $\hat{\mathcal{A}}_n = \hat{\mathcal{A}}_n \cup \{r\}$ and

$$\beta = \text{argmin}_{\beta: \beta_j = 0, \beta_{jk} = 0, j \in \hat{\mathcal{A}}_n^c, (j,k) \in \hat{\mathcal{A}}_n^c} \| Y - X'\beta - Z'B \|_n^2.$$

- Hao and Zhang [2014]

Strong hierarchical lasso (Choi et al. [2010]) I

- Reparametrize the coefficients for the interaction terms.

$$\beta_{jk} = \alpha_{jk}\beta_j\beta_k,\ j < k$$

- Rewrite the model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{j<k} \alpha_{jk}\beta_j\beta_k X_j X_k$$

where $\alpha = (\alpha_{jk})$

- Find β and α as

$$\underset{\beta,\alpha}{\operatorname{argmin}} \| Y - X'\beta - Z'\alpha\beta \otimes \beta\|_n^2 + \lambda_\beta\|\beta\|_1 + \lambda_\alpha\|\alpha\|_1 \qquad (1)$$

Strong hierarchical lasso (Choi et al. [2010]) II

- Whenever the coefficient for either X_j or X_k, that is, β_j, β_k, is equal to zero, the coefficient for the interaction term $X_j X_k$ is automatically set to zero.

Strong hierarchical lasso (Choi et al. [2010]) III

- The estimator can be obtained by minimzing the objective function by iteratively minimizing β_k and α_{jk}.
- Computation is very demanding because
  - the number of parameters is large, and
  - the objective function is not convex.
- Almost impossible to be scaled-up for very high dimensional models.

Relaxed Lasso for interactions with heredity (Bien et al. [2013]) I

- Relaxed lasso for strong heredity
  - Find β and B as

$$\underset{\beta \in R^p, B \in R^{p \times p}}{\operatorname{argmin}} \| Y - X'\beta^+ + X'\beta^- - Z'B \|_2^n + \lambda 1'(\beta^+ + \beta^-) + \frac{\lambda}{2}\|B\|_1$$

s.t. $B = B'$, $\|B_j\|_1 \leq \beta_j^+ + \beta_j^-$, $\beta_j^+ \geq 0, \beta_j^- \geq 0$
for $j = 1, \cdots, p$

Relaxed Lasso for interactions with heredity (Bien et al. [2013]) II

- Relaxed lasso for weak heredity
  - Find β and B as

$$\underset{\beta \in R^p, B \in R^{p \times p}}{\operatorname{argmin}} \; \| Y - X'\beta^+ + X'\beta^- - Z'B \|_2^n + \lambda 1'(\beta^+ + \beta^-) + \frac{\lambda}{2} \|B\|_1$$

s.t. $\|B_j\|_1 \leq \beta_j^+ + \beta_j^-, \beta_j^+ \geq 0, \beta_j^- \geq 0$
for $j = 1, \cdots, p$

Relaxed Lasso for interactions with heredity (Bien et al. [2013]) III

- Bien et al. [2013] suggest a generalized gradient descent algorithm to calculate the solution.
- It requires few thousands iterations until convergence while p^2 many computations is required in each iteration.
- As Bien et al. [2013] noted, the algorithm can be applicable only when p is less than few hundreds.

LARS algorithm with heredity (Kim and Kim [2017]) I

- Initialization: $\hat{\mathcal{A}}_n = \emptyset \beta = 0, \gamma > 0$
- Repeat
  - $R = Y - X^{'}\beta - Z^{'}B$
  - Choose $j = \text{argmax}_{k \in \{1,\ldots,p\} - \hat{\mathcal{A}}_n^{(1)}} |\text{cov}_n(R, X_k)|$
  - Choose $(l, h) = \text{argmax}_{(j,k) \in \mathcal{I}_{\hat{\mathcal{A}}_n} - \hat{\mathcal{A}}_n^{(2)}} |\text{cov}_n(R, X_j X_k)|$
  - Let $r = j$ if $|\text{cov}_n(R, X_j)| \geq |\text{cov}_n(R, X_l X_h)|$ or let $r = (l, h)$ otherwise.
  - $\hat{\mathcal{A}}_n = \hat{\mathcal{A}}_n \cup \{r\}$ and

$$\beta_r = \beta_r + \gamma \text{sign}\{\text{cov}_n(R, X_r)\} \text{ if } r \in \{1, \ldots, p\}$$

or

$$\beta_r = \beta_r + \gamma \text{sign}\{\text{cov}_n(R, Z_r)\} \text{ if } r \in \{1, \ldots, p\}^2.$$

3. Interaction Screening under the Gaussian assumption

Aim

- We want to select interactions without heredity.
- That is, we want to detect important interactions even when the corresponding main effects are not significant.
- Under the assumption that the distribution of $X = (X_1, \ldots, X_p)$ is Gaussian, we propose **a measure between Y and X_j which is large when some interactions involving X_j are significant**.

Some properties of Gaussian distribution I

- $(X_1, \ldots, X_p)$ are Gaussianly distributed with $\mathrm{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$.
- Let $r_{jk} = \mathrm{cov}(X_j, X_k)$.
- Then, we have
  - $\mathrm{E}(X_l | X_j) = r_{jl} X_j$
  - $\mathrm{E}(X_l X_h | X_j) = (r_{lh} - r_{jl} r_{jh}) + r_{jl} r_{jh} X_j^2$.

Some properties of Gaussian distribution II

- Hence, we have

$$\mathrm{E}(Y|X_j) = \alpha_j + \mu_{1j}X_j + \mu_{2j}X_j^2$$

where

$$\mu_{1j} = \sum_{l=1}^{p} \beta_l^* r_{lj}$$

and

$$\mu_{2j} = \sum_{l \le h} \beta_{lh}^* r_{jl} r_{jh}.$$

Our proposal: SIS for interaction I

- Let $\hat{\mu}_{1j}$ and $\hat{\mu}_{2j}$ be the minimizer of

$$\sum_{i=1}^{n} \left(Y_i - \alpha_j - \mu_{1j} X_{ij} - \mu_{2j} X_{ij}^2 \right)^2 .$$

- Choose two thresholds τ_{1n} and τ_{2n}.
- Let

$$\hat{\mathcal{A}}_{n1} = \{ j : |\hat{\mu}_{1j}| > \tau_{1n} \}$$

$$\hat{\mathcal{A}}_{n2} = \{ j : |\hat{\mu}_{2j}| > \tau_{2n} \}.$$

- Select main effects in $\hat{\mathcal{A}}_{n1}$.
- For screening interaction, we apply the standard SIS to $X_j X_k, k = 1, \ldots, p$ only for $j \in \hat{\mathcal{A}}_{n2}$.

Our proposal: SIS for interaction II

- Let

$$\mathcal{A}_1^* = \{j : \beta_j^* \neq 0\}$$

and

$$\mathcal{A}_2^* = \{j : \beta_{jk}^* \neq 0 \text{ for some } k\}.$$

- Essentially, SIS-Interaction estimates (does screening) $\mathcal{A}_1^*$ and $\mathcal{A}_2^*$ by $\hat{\mathcal{A}}_{n1}$ and $\hat{\mathcal{A}}_{n2}$, respectively.
- Does SIS-Interaction have screening property?

Revists SIS I

- Let $w_j = \mathrm{cov}_n(Y, X_j)$, where $\mathrm{cov}_n(Y, X_j)$ is the empirical covariance between Y and X_j with the sample size n.
- Let $\hat{\mathcal{A}}_n = \{j : |w_j| > \tau_n\}$ for some threshold $\tau_n > 0$.
- Under regularity conditions, $\hat{A}_n$ has a screening property provided
  (1) $|\mathrm{cov}(Y, X_j)|$ is large for $\beta_j^* \neq 0$
  (2) $|\mathrm{cov}(Y, X_j)|$ is small for most of $\beta_j^* = 0$.

Revists SIS II

- Note that

$$\mathrm{cov}(Y, X_j) = \sum_{l=1}^{p} \beta_l^* r_{lj}$$

  which is equal to μ_{1j}.
- We can write

$$\mu_{1j} = \beta_j^* \left(1 + \frac{\sum_{l \neq j} \beta_l^* r_{lj}}{\beta_j^*} \right).$$

Revists SIS III

- Hence, the SIS conditions (1) and (2) hold when
  (1*) $|\beta_j^*|$ is large and $|\sum_{l \neq j} \beta_l^* r_{lj}|$ is small when $\beta_j^* \neq 0$
  (2*) $|\sum_{l \neq j} \beta_l^* r_{lj}|$ is small for most j with $\beta_j^* = 0$.
- Condition (2*) holds when the largest eigenvalue of the design matrix is not too large (e.g. polynomial order of n).

About the conditions for SIS-Interaction I

- Let $\mathcal{A}_2^* = \{j : \beta_{jk}^* \neq 0 \text{ for some } k\}$.
- SIS-Interaction assumes not only the conditions (1) and (2) of the standard SIS but also assumes that
  - (3) $|\mu_{2j}|$ is large for $j \in \mathcal{A}_2^*$
  - (4) $|\mu_{2j}|$ is small for most $j \notin \mathcal{A}_2^*$
- Let

$$\xi_j = \sum_l \beta_{lj}^* r_{lj}$$

$$\tilde{\mu}_{2j} = \sum_{l \leq h, (l,h) \neq j} \beta_{lh}^* r_{lj} r_{hj}.$$

About the conditions for SIS-Interaction II

- Note that

$$\mu_{2j} = \xi_j \left(1 + \frac{\tilde{\mu}_{2j}}{\xi_j}\right) \text{ for } j \in \mathcal{A}_2^*$$

and

$$\mu_{2j} = \tilde{\mu}_{2j} \text{ for } j \notin \mathcal{A}_2^*.$$

- Thus, the conditions (3) and (4) for SIS-Interaction can be rephrased as

(3*) $|\xi_j|$ is large and $|\tilde{\mu}_{2j}|$ is small for $j \in \mathcal{A}_2^*$
(4*) $|\tilde{\mu}_{2j}|$ is small for most $j \notin \mathcal{A}_2^*$.

- Hence the conditions (3) and (4) are similar to the conditions (1) and (2).

Remark about the condition (3*)

- There are models where $\xi_j = 0$ even when $j \in \mathcal{A}_2^*$.
- Example (so called XOR model)
  - $Y = \beta_{12}^* X_1 X_2 + \epsilon$
  - X_1 and X_2 are independnent.
  - $\xi_1 = \beta_{12}^* r_{12} = 0$ since $r_{12} = 0$.
- It is not suprising that XOR model cannot be found based only on the marginal relations between response and covariates,
- Correlations between covariates are indispensable for SIS-Interaction.

Remark about the condition (4*)

- (4*) would hold when the largest eigenvalue of **the design matrix of the main effects** is not too large!
- Under construction though...

Computational consideration

- When p is large, it would be computational demanding to get $\hat{\mu}_{1j}$ and $\hat{\mu}_{2j}$ for $j = 1, \ldots, p$ which requires p many times matrix inversion of the 3×3 matrix to minimize

$$\sum_{i=1}^{n} \left(Y_i - \alpha_j - \mu_{1j} X_{ij} - \mu_{2j} X_{ij}^2 \right)^2 . \tag{2}$$

- Note that $\mathrm{E}(YX_j) = \mu_{1j}^*$. Hence, we can estimate μ_{1j} by its empirical counter part $\mathrm{cov}_n(Y, X_j)$.
- If we plug the estimate of μ_{1j} in (2), we can find the least square estimator of μ_{2j} by simple algebra without matrix inversion.

Application to LARS (and Forward selection)

- At each iteration, we calculate $\hat{\mu}_{2j}$ between the residuals R and X_j and let $\hat{j} = \text{argmax}|\hat{\mu}_{2j}|$.
- Choose the main effect which maximizes $|\text{cov}_n(R, X_j)|$
- Choose the interaction that maximizes $|\text{cov}_n(R, X_{\hat{j}}X_k)|$.
- Add either the main effect or the interaction whichever has larger absolute covariance with the residual.
- Update the selected regression coefficient accordingly.

Extension to multivariate t-distribution I

- Suppose $X = (X_1, \ldots, X_p)'$ follows the multivariate t-distribution with parameters (ν, Σ), where the density of the multivariate t-distribution with parameters (ν, Σ) is given as

$$f(\mathbf{x}) \propto \left(1 + \frac{1}{\nu}\mathbf{x}'\Sigma^{-1}\mathbf{x}\right)^{-\frac{\nu+p}{2}}$$

where $f(t : \nu)$ is the density of the t-distribution with degree of freedom ν.

Extension to multivariate t-distribution II

- Then, we have

$$\mathrm{E}(Y|X_j) = \alpha_j^{(t)} + \mu_{1j}^{(t)} X_j + \mu_{2j}^{(t)} X_j^2,$$

where

$$\mu_{1j}^{(t)} = \mu_{1j}$$

and

$$\mu_{2j}^{(t)} = \gamma_0 + \frac{\nu}{\nu + 1} \mu_{2j}$$

for some constant γ_0.

Extension to multivariate t-distribution III

- When $p \to \infty$ and the largest eignevalue of Σ is not too large (compared to p), it can be shown that

$$\bar{\mu}_2^{(t)} = \frac{\sum_{j=1}^{p} \mu_{2j}^{(t)}}{p} \to \gamma_0.$$

- Hence, asymptotically,

$$\mu_{2j}^{(t)} - \bar{\mu}_2^{(t)} \approx \frac{\nu}{\nu + 1} \mu_{2j}.$$

- We can apply the SIS-I with replacing $\hat{\mu}_{2j}$ by $\hat{\mu}_{2j}^{(t)} - \hat{\bar{\mu}}_2^{(t)}$.

SIS-interaction for t-distribution

- Let $\hat{\mu}_{1j}$ and $\hat{\mu}_{2j}$ be the minimizer of

$$\sum_{i=1}^{n} \left(Y_i - \alpha_j - \mu_{1j} X_{ij} - \mu_{2j} X_{ij}^2 \right)^2 .$$

- Let $\hat{\mu}_{2j} = \hat{\mu}_{2j} - \sum_{k=1}^{p} \hat{\mu}_{2k}/p$
- Let

$$\hat{\mathcal{A}}_{n1} = \{ j : |\hat{\mu}_{1j}| > \tau_{1n} \}$$

$$\hat{\mathcal{A}}_{n2} = \{ j : |\hat{\mu}_{2j}| > \tau_{2n} \}.$$

- Select main effects in $\hat{\mathcal{A}}_{n1}$.
- For screening interaction, we apply the standard SIS to $X_j X_k, k = 1, \ldots, p$ only for $j \in \hat{\mathcal{A}}_{n2}$.

4. Simulation

Simulation set-up

- Compare LARS with Forward selection
- We select the optimal subset using the high-dimensional BIC [Wang et al., 2013]
- We set $n = 100$.
- We generate main effects from Uniform(-10,10) and interactions from Uniform(-5,5).
- We set 5 active intereactions (and corresponding main effects for heredity).

With heredity

	sensitivity			specificity		
	p=5	p=50	p=100	p=5	p=50	p=100
LARS-H	0.9988	0.7708	0.6206	0.4345	0.9526	0.9904
FRWD-H	0.9953	0.5390	0.3936	0.7010	0.9969	0.9995
LASSO	0.9987	0.9896	0.9807	0.3678	0.9477	0.9861

Table : Sensitivity and specificity with heredity

Without heredity

	sensitivity			specificity		
	p=5	p=50	p=100	p=5	p=50	p=100
LARS-I	1	0.9375	0.7200	0.6050	0.9643	0.9910
FWRD-I	0.9975	0.9228	0.5557	0.9467	0.9982	0.9994
LARS-H	1	0.6260	0.4257	0.3160	0.9508	0.9913
FWRD-H	0.9975	0.2753	0.1628	0.6560	0.9979	0.9996

Table : Sensitivity and specificity without heredity

Thank you!

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

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

R. Cox, D. Interaction. *International Statistical Review*, 52(1): 1–24, 1984.

B. Efron, T. Hastie, I. Johnston, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–451, 2004.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

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

Woosung Kim and Yongdai Kim. Modified lars algorithm for interaction selection with heredity. 2017.

Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505, 2013.