Detection of Cooperative Interactions in Logistic Regression Models

"Easton" Li Xu

Texas A&M University

(Joint work with Xiaoning Qian, Tie Liu (Texas A&M), Shuguang Cui (UC Davis))

August 2016

Linear Regression Models

- d input variables $X_1, X_2, \ldots, X_d$ and an output variable Y

Linear Regression Models

- d input variables $X_1, X_2, \ldots, X_d$ and an output variable Y
- Linear regression model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

Linear Regression Models

- d input variables $X_1, X_2, \ldots, X_d$ and an output variable Y
- Linear regression model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

- Training samples: observations of $(Y, X_1, X_2, \ldots, X_d)$

$$\begin{aligned} \text{Sample } 1 &: (y^{(1)}, x_1^{(1)}, x_2^{(1)}, \ldots, x_d^{(1)}) \\ &\vdots \\ \text{Sample } n &: (y^{(n)}, x_1^{(n)}, x_2^{(n)}, \ldots, x_d^{(n)}) \end{aligned}$$

Linear Regression Models

- d input variables $X_1, X_2, \ldots, X_d$ and an output variable Y
- Linear regression model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

- Training samples: observations of $(Y, X_1, X_2, \ldots, X_d)$

$$\text{Sample 1} : (y^{(1)}, x_1^{(1)}, x_2^{(1)}, \ldots, x_d^{(1)})$$
$$\vdots$$
$$\text{Sample } n : (y^{(n)}, x_1^{(n)}, x_2^{(n)}, \ldots, x_d^{(n)})$$

- Parameter estimation:

$$(\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_d) = \arg\min_{\beta_1, \beta_2, \ldots, \beta_d} \frac{1}{n} \sum_{t=1}^{n} \left| y^{(t)} - (\beta_1 x_1^{(t)} + \ldots + \beta_d x_d^{(t)}) \right|^2$$

Linear Regression Models

- d input variables $X_1, X_2, \ldots, X_d$ and an output variable Y
- Linear regression model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

- Training samples: observations of $(Y, X_1, X_2, \ldots, X_d)$

$$\text{Sample } 1 : (y^{(1)}, x_1^{(1)}, x_2^{(1)}, \ldots, x_d^{(1)})$$

$$\vdots$$

$$\text{Sample } n : (y^{(n)}, x_1^{(n)}, x_2^{(n)}, \ldots, x_d^{(n)})$$

- Parameter estimation:

$$(\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_d) = \underset{\beta_1, \beta_2, \ldots, \beta_d}{\arg\min} \frac{1}{n} \sum_{t=1}^{n} \left| y^{(t)} - (\beta_1 x_1^{(t)} + \ldots + \beta_d x_d^{(t)}) \right|^2$$

- Test sample: $(x_1^{\text{test}}, x_2^{\text{test}}, \ldots, x_d^{\text{test}})$
- Prediction:

$$y^{\text{test}} = \tilde{\beta}_1 x_1^{\text{test}} + \cdots + \tilde{\beta}_d x_d^{\text{test}}$$

Logistic Regression Models

- Linear regression models

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d \triangleq \boldsymbol{\beta} \cdot \mathbf{X}$$

Logistic Regression Models

- Linear regression models

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d \triangleq \boldsymbol{\beta} \cdot \mathbf{X}$$

- Logistic regression models

$$\begin{aligned} \Pr(Y = +1|\mathbf{X}) &= \boldsymbol{\beta} \cdot \mathbf{X} \\ \Pr(Y = -1|\mathbf{X}) &= 1 - \Pr(Y = +1|\mathbf{X}) \end{aligned}$$

Logistic Regression Models

- Linear regression models

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d \triangleq \boldsymbol{\beta} \cdot \mathbf{X}$$

- Logistic regression models

$$\begin{aligned} \Pr(Y = +1|\mathbf{X}) &= \boldsymbol{\beta} \cdot \mathbf{X} \\ \Pr(Y = -1|\mathbf{X}) &= 1 - \Pr(Y = +1|\mathbf{X}) \end{aligned}$$

$$\Downarrow \; \sigma(x) := 1/(1 + e^{-x}) \in [0, 1]$$

$$\begin{aligned} \Pr(Y = +1|\mathbf{X}) &= \sigma(\boldsymbol{\beta} \cdot \mathbf{X}) \\ \Pr(Y = -1|\mathbf{X}) &= 1 - \Pr(Y = +1|\mathbf{X}) \end{aligned}$$

Logistic Regression Models

- Linear regression models

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d \triangleq \boldsymbol{\beta} \cdot \mathbf{X}$$

- Logistic regression models

$$\begin{aligned} \Pr(Y = +1|\mathbf{X}) &= \boldsymbol{\beta} \cdot \mathbf{X} \\ \Pr(Y = -1|\mathbf{X}) &= 1 - \Pr(Y = +1|\mathbf{X}) \end{aligned}$$

$$\Downarrow \; \sigma(x) := 1/(1 + e^{-x}) \in [0, 1]$$

$$\begin{aligned} \Pr(Y = +1|\mathbf{X}) &= \sigma(\boldsymbol{\beta} \cdot \mathbf{X}) \\ \Pr(Y = -1|\mathbf{X}) &= 1 - \Pr(Y = +1|\mathbf{X}) \end{aligned}$$

Individual Effects and Pairwise Interactions

Logistic regression model with individual effects and pairwise interactions

$$\Pr(Y = +1|\mathbf{X}) = \sigma(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d$$

Individual Effects and Pairwise Interactions

Logistic regression model with individual effects and pairwise interactions

$$\Pr(Y = +1|\mathbf{X}) = \sigma(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d \\ + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \cdots + \beta_{d-1,d} X_{d-1} X_d)$$

Individual Effects and Pairwise Interactions

Logistic regression model with individual effects and pairwise interactions

$$\Pr(Y = +1|\mathbf{X}) = \sigma(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \cdots + \beta_{d-1,d} X_{d-1} X_d)$$

- $\beta_i \neq 0$: X_i has an individual effect.
- $\beta_i = 0$: X_i has no individual effect.
- $\beta_{i,j} \neq 0$: X_i and X_j has a pairwise interaction.
- $\beta_{i,j} = 0$: X_i and X_j has no pairwise interaction.

System Model

- $X_1, X_2, \ldots, X_d$ are independent variables with $\Pr\{X_i = +1\} = \Pr\{X_i = -1\} = 1/2$, for $i = 1, 2, \ldots, d$.
- Y is a binary outcome variable

$$
\begin{aligned}
\Pr\{Y = +1 | X_1, X_2, \ldots, X_d\} &= \sigma\Big(\sum_{i=1}^{d} \beta_i X_i + \sum_{1 \le i < j \le d} \beta_{i,j} X_i X_j\Big) \\
\Pr\{Y = -1 | X_1, X_2, \ldots, X_d\} &= 1 - \Pr\{Y = +1 | X_1, X_2, \ldots, X_d\} \\
&= \sigma\Big(-\sum_{i=1}^{d} \beta_i X_i - \sum_{1 \le i < j \le d} \beta_{i,j} X_i X_j\Big)
\end{aligned}
$$

System Model

- $X_1, X_2, \ldots, X_d$ are independent variables with $\Pr\{X_i = +1\} = \Pr\{X_i = -1\} = 1/2$, for $i = 1, 2, \ldots, d$.
- Y is a binary outcome variable

$$\begin{aligned}
\Pr\{Y = +1 | X_1, X_2, \ldots, X_d\} &= \sigma\Big(\sum_{i=1}^{d} \beta_i X_i + \sum_{1 \le i < j \le d} \beta_{i,j} X_i X_j\Big) \\
\Pr\{Y = -1 | X_1, X_2, \ldots, X_d\} &= 1 - \Pr\{Y = +1 | X_1, X_2, \ldots, X_d\} \\
&= \sigma\Big(-\sum_{i=1}^{d} \beta_i X_i - \sum_{1 \le i < j \le d} \beta_{i,j} X_i X_j\Big)
\end{aligned}$$

Target:

Detect all individual effects and pairwise interactions in logistic regression models from a limited number of samples.

Motivation 1: Detection of the Graph Underlying an Ising Model [Bresler (2015)]

- Ising models on a graph $G = (V, E)$ with $|V| = d$:

$$p(X_1, X_2, \ldots, X_d) = \exp\left\{\sum_{i \in V} \beta_i X_i + \sum_{\{i,j\} \in E} \beta_{i,j} X_i X_j - \Phi(\beta)\right\}$$

- parameter vector: $\beta = \{\beta_i\}_{i \in V} \cup \{\beta_{i,j}\}_{\{i,j\} \in E}$
- normalizing constant: $\Phi(\beta)$
- the maximum degree of nodes is p (constant)
- $|\beta_i| \leq h$ and $\lambda \leq |\beta_{i,j}| \leq \mu$.

Motivation 1: Detection of the Graph Underlying an Ising Model [Bresler (2015)] (Continued)

Theorem (Bresler 2015)

Let $\delta = \frac{1}{2}e^{-2(\mu p + h)}$, $\tau^* = \frac{\lambda^2 \delta^{4p+1}}{16 p \mu}$, $\epsilon^* = \frac{\tau^*}{2}$, $\ell^* = \frac{8}{(\tau^*)^2}$. *Suppose we observe* n *samples with*

$$n \geq \frac{144(\ell^* + 3)}{(\epsilon^*)^2 \delta^{2\ell^*}} \log \frac{d}{\zeta}.$$

Then with probability at least $1 - \zeta$, *there exists an algorithm to detect the structure of* G *running in polynomial time* $O(\ell^* dn)$.

Motivation 2: Chow-Liu Tree [Chow & Liu (1968)]

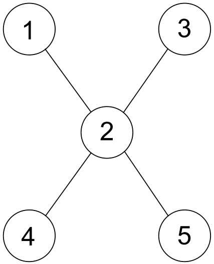
Chow-Liu representation:

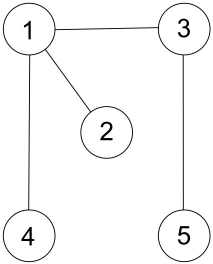
$$
\begin{aligned}
& p(X_1, X_2, X_3, X_4, X_5) \\
= \; & p(X_1) \cdot p(X_2|X_1) \cdot p(X_3|X_1, X_2) \cdot p(X_4|X_1, X_2, X_3) \cdot p(X_5|X_1, X_2, X_3, X_4) \\
\approx \; & p(X_1) \cdot p(X_2|X_1) \cdot p(X_3|X_2) \qquad \cdot p(X_4|X_2) \qquad \cdot p(X_5|X_2) \\
& \text{(first-order product approximation)} \\
= \; & p'(X_1, X_2, X_3, X_4, X_5)
\end{aligned}
$$

Target: Find p' to minimize the Kullback-Leibler distance $D(p||p')$ between p and p'.

Motivation 2: Chow-Liu Tree [Chow & Liu (1968)] (Continued)

Dependency Relationship





$$p(X_1, X_2, X_3, X_4, X_5) \approx p(X_1)p(X_2|X_1)p(X_3|X_2)p(X_4|X_2)p(X_5|X_2)$$

$$p(X_1, X_2, X_3, X_4, X_5) \approx p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_1)p(X_5|X_3)$$

Motivation 2: Chow-Liu Tree [Chow & Liu (1968)] (Continued)

Chow-Liu Algorithm:

- Construct a weighted complete graph $G = (V, E)$ with $V = \{v_1, v_2, \ldots, v_d\}$.
- The weight $w(v_i, v_j)$ of edge (v_i, v_j) is assigned to be $I(X_i; X_j)$.
- Find a maximum spanning tree T of G (by Kruskal's algorithm or Prim's algorithm).
- Set an arbitrarily node v to be the root of T, then rank the other nodes by their depths.

Our Work

- Model all individual effects and pairwise interaction by a so-called interaction graph.

Our Work

- Model all individual effects and pairwise interaction by a so-called interaction graph.
- Establish an algorithm with a similar style as Chow-Liu algorithm to detect the structure of the interaction graph from a limited number of samples.

Our Work

- Model all individual effects and pairwise interaction by a so-called interaction graph.
- Establish an algorithm with a similar style as Chow-Liu algorithm to detect the structure of the interaction graph from a limited number of samples.
- No assumption of the maximum degree of nodes.

Our Work

- Model all individual effects and pairwise interaction by a so-called interaction graph.
- Establish an algorithm with a similar style as Chow-Liu algorithm to detect the structure of the interaction graph from a limited number of samples.
- No assumption of the maximum degree of nodes.
- Sample complexity and running time are both polynomial functions of the number of features.

Model with only Pairwise Interactions

- **Assumption:**
  No individual effects ($\beta_i = 0$ for $1 \leq i \leq d$).

- **For example:**
  - 5 variables X_1, X_2, X_3, X_4, X_5
  - $\beta_{1,2}, \beta_{2,3}, \beta_{2,4}, \beta_{2,5} \neq 0$ and other $\beta_{i,j} = 0$

$$\begin{aligned}
\Pr\{Y = +1 | X_1, X_2, X_3, X_4, X_5\} &= \sigma(\beta_{1,2}X_1X_2 + \beta_{2,3}X_2X_3 \\
&\quad + \beta_{2,4}X_2X_4 + \beta_{2,5}X_2X_5) \\
\Pr\{Y = -1 | X_1, X_2, X_3, X_4, X_5\} &= 1 - \Pr\{Y = +1 | X_1, X_2, X_3, X_4, X_5\}
\end{aligned}$$

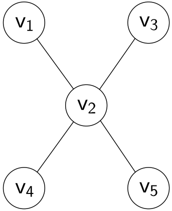
Interaction Graph

Interaction graph: Let $G = (V, E)$ be the interaction graph with $V = \{v_1, v_2, \dots, v_d\}$, and the edge $(v_i, v_j) \in E$ if and only if the coefficient $\beta_{i,j}$ corresponding to X_i and X_j is nonzero.

For example:

$$\begin{aligned}
&\Pr\{Y = +1 | X_1, X_2, X_3, X_4, X_5\} \\
&\quad = \sigma(\beta_{1,2}X_1X_2 + \beta_{2,3}X_2X_3 \\
&\qquad + \beta_{2,4}X_2X_4 + \beta_{2,5}X_2X_5)
\end{aligned}$$

$$\begin{aligned}
&\Pr\{Y = -1 | X_1, X_2, X_3, X_4, X_5\} \\
&\quad = 1 - \Pr\{Y = +1 | X_1, X_2, X_3, X_4, X_5\}
\end{aligned}$$



$\beta_{1,2}, \beta_{2,3}, \beta_{2,4}, \beta_{2,5} \neq 0$

Assumption, Difficulty & Target

- **Assumption:**

  The interaction graph $G = (V, E)$ is acyclic.

  - When the model contains at most two interactions, G is always acyclic.
  - When the number of interactions is far less than the number of features, G is acyclic with a high probability.
  - The model contains at most $d - 1$ interactions.

- **Difficulty:**

  We don't know which edges this graph has.

- **Target:**

  Detect the structure of the interaction graph from a limited number of samples.

Construction of a Weighted Complete Graph

Construction:

Construct a weighted complete graph $G' = (V', E')$ by

- $V' = (v'_1, v'_2, \dots, v'_d)$
- The weight of any edge $(v'_i, v'_j) \in E'$ is

$$w_{\{i,j\}} = \big| \Pr\{Y = +1 | X_i = +1, X_j = +1\} - \Pr\{Y = -1 | X_i = +1, X_j = +1\} \big|.$$

Structure Detection of the Interaction Graph (Case 1)

- Case 1: The third-order joint probability $p(X_i, X_j, Y)$ is known.
- $w_{\{i,j\}}$ can be calculated from the third-order joint distribution of X_i, X_j, Y

$$
\begin{aligned}
& w_{\{i,j\}} \\
= & |\Pr\{Y = +1|X_i = +1, X_j = +1\} - \Pr\{Y = -1|X_i = +1, X_j = +1\}| \\
= & |8 \Pr\{X_i = +1, X_j = +1, Y = +1\} - 1|
\end{aligned}
$$

Theorem on Detection (Case 1)

Theorem

Let $T = (V', E_T)$ be a maximum spanning tree of G'. Then

$$(v_i, v_j) \in E \text{ if and only if } (v_i', v_j') \in E_T \text{ and } w_{\{i,j\}} > 0.$$

Theorem on Detection (Case 1)

Theorem

Let $T = (V', E_T)$ *be a maximum spanning tree of* G'. *Then*

$$(v_i, v_j) \in E \textit{ if and only if } (v_i', v_j') \in E_T \textit{ and } w_{\{i,j\}} > 0.$$

edges in the interaction graph

$\Updownarrow$

non-zero weighted edges in the maximum spanning tree

Detection Algorithm (Case 1)

Algorithm (Detecting the interaction graph)

- *Construct a weighted graph* $G' = (V', E')$ *with* $V' = \{v'_1, v'_2, \ldots, v'_d\}$.
- *The weight* $w_{\{i,j\}}$ *of edge* (v'_i, v'_j) *is assigned to be* $|\Pr\{Y = +1 | X_i = +1, X_j = +1\} - \Pr\{Y = -1 | X_i = +1, X_j = +1\}|$.
- *Find a maximum spanning tree* $T' = (V', E_T)$ *of* G' *(by Kruskal's algorithm or Prim's algorithm).*
- *Then the set of the edges in* G *is* $\{(v_i, v_j) : (v'_i, v'_j) \in E_T \text{ and } w_{\{i,j\}} > 0\}$.

Detection Algorithm (Case 1)

Algorithm (Detecting the interaction graph)

- *Construct a weighted graph* $G' = (V', E')$ *with* $V' = \{v'_1, v'_2, \ldots, v'_d\}$.
- *The weight* $w_{\{i,j\}}$ *of edge* (v'_i, v'_j) *is assigned to be* $|\Pr\{Y = +1 | X_i = +1, X_j = +1\} - \Pr\{Y = -1 | X_i = +1, X_j = +1\}|$.
- *Find a maximum spanning tree* $T' = (V', E_T)$ *of* G' *(by Kruskal's algorithm or Prim's algorithm).*
- *Then the set of the edges in* G *is* $\{(v_i, v_j) : (v'_i, v'_j) \in E_T \text{ and } w_{\{i,j\}} > 0\}$.

The algorithm is executed in polynomial time $O(d^2)$.

Structure Detection of the Interaction Graph (Case 2)

- Case 2:
  - The third-order joint probability $p(X_i, X_j, Y)$ is unknown.
  - Any non-zero parameter $\beta_{i,j}$ satisfies that

$$\lambda \leq |\beta_{i,j}| \leq \mu.$$

- Weight Assignment: With n samples $(Y^{(t)}, X_1^{(t)}, X_2^{(t)}, \ldots, X_d^{(t)})$ for $1 \leq t \leq n$, we estimate

$$w_{\{i,j\}} = |8\,\mathrm{Pr}\{X_i = +1, X_j = +1, Y = +1\} - 1|$$

by

$$\hat{w}_{\{i,j\}} = \left| \frac{8}{n} \sum_{t=1}^{n} \mathbf{1}_{(X_i^{(t)}, X_j^{(t)}, Y^{(t)}) = (+1,+1,+1)} - 1 \right|.$$

Theorem on Detection (Case 2)

Let

$$\gamma = \sqrt{\frac{2}{\pi d}} \left[\sigma(\lambda + 3\mu) - \sigma(-\lambda + 3\mu)\right].$$

Theorem

Assume for $1 \leq i < j \leq d$,

$$|\hat{w}_{\{i,j\}} - w_{\{i,j\}}| < \gamma/2.$$

Let $T = (V', E_T)$ *be a maximum spanning tree of* G'. *Then*

$$(v_i, v_j) \in E \text{ if and only if } (v_i', v_j') \in E_T \text{ and } \hat{w}_{\{i,j\}} > \gamma/2.$$

Theorem on Detection (Case 2)

Let

$$\gamma = \sqrt{\frac{2}{\pi d}} \left[\sigma(\lambda + 3\mu) - \sigma(-\lambda + 3\mu) \right].$$

Theorem

Assume for $1 \leq i < j \leq d$,

$$\left| \hat{w}_{\{i,j\}} - w_{\{i,j\}} \right| < \gamma/2.$$

Let $T = (V', E_T)$ *be a maximum spanning tree of* G'. *Then*

$$(v_i, v_j) \in E \text{ if and only if } (v_i', v_j') \in E_T \text{ and } \hat{w}_{\{i,j\}} > \gamma/2.$$

edges in the interaction graph

$\Updownarrow$

large weighted edges in the maximum spanning tree

Detection Algorithm (Case 2)

Algorithm (Detecting the interaction graph)

- *Construct a weighted graph* $G' = (V', E')$ *with* $V' = \{v'_1, v'_2, \ldots, v'_d\}$.
- *The weight* $w_{\{i,j\}}$ *of edge* (v'_i, v'_j) *is assigned to be*

$$\left| \frac{8}{n} \sum_{t=1}^{n} \mathbf{1}_{(X_i[t], X_j[t], Y[t]) = (+1,+1,+1)} - 1 \right|.$$

- *Find a maximum spanning tree* $T' = (V', E_T)$ *of* G' *(by Kruskal's algorithm or Prim's algorithm).*
- *Then the set of the edges in* G *is* $\{(v_i, v_j) : (v'_i, v'_j) \in E_T \text{ and } w_{\{i,j\}} > \gamma/2\}$.

The algorithm is executed in polynomial time $O(nd^2)$.

Sample Complexity (Case 2)

Theorem

Fix $0 < \epsilon < 1$ and let n be a positive integer such that

$$n \geq \frac{128}{\gamma^2} \log \frac{d^2}{\epsilon} = \frac{64\pi d}{[\sigma(\lambda + 3\mu) - \sigma(-\lambda + 3\mu)]^2} \log \frac{d^2}{\epsilon}. \quad (1)$$

Then with probability at least $1 - \epsilon$, the algorithm can successfully detect the graph G from n i.i.d. samples of $(X_1, X_2, \ldots, X_d, Y)$.

The order of sample complexity: $\Theta\left(d \log \frac{d}{\epsilon}\right)$
Running time: $O(d^3 \log \frac{d}{\epsilon})$

Models with both Individual Effects and Pairwise Interactions

- **For example:**
  - 4 variables X_1, X_2, X_3, X_4
  - $\beta_2, \beta_{1,2}, \beta_{2,3}, \beta_{2,4} \neq 0$ and other $\beta_i, \beta_{i,j} = 0$

$$\begin{aligned}\Pr\{Y = +1 | X_1, X_2, X_3, X_4\} &= \sigma(\beta_2 X_2 + \beta_{1,2} X_1 X_2 \\ &\quad + \beta_{2,3} X_2 X_3 + \beta_{2,4} X_2 X_4) \\ \Pr\{Y = -1 | X_1, X_2, X_3, X_4\} &= 1 - \Pr\{Y = +1 | X_1, X_2, X_3, X_4\}\end{aligned}$$

Extended Interaction Graph

For extended interaction graph $G = (V, E)$,

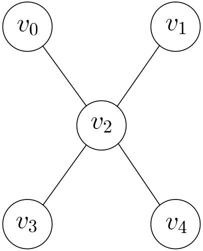
- $V = \{v_0(\text{virtual vertex}), v_1, v_2, \ldots, v_d\}$
- $(v_0, v_i) \in E$ if and only if X_i has an individual effect
- $(v_i, v_j) \in E$ if and only if X_i and X_j have a cooperative interaction

With the help of the virtual vertex v_0, G can capture all individual effects and pairwise interactions.

For example:

$$\begin{aligned}
&\Pr\{Y = +1 | X_1, X_2, X_3, X_4\} \\
&\quad = \sigma(\beta_2 X_2 + \beta_{1,2} X_1 X_2 \\
&\qquad + \beta_{2,3} X_2 X_3 + \beta_{2,4} X_2 X_4)
\end{aligned}$$

$$\begin{aligned}
&\Pr\{Y = -1 | X_1, X_2, X_3, X_4\} \\
&\quad = 1 - \Pr\{Y = +1 | X_1, X_2, X_3, X_4\}
\end{aligned}$$



$\beta_2, \beta_{1,2}, \beta_{2,3}, \beta_{2,4} \neq 0$

Auxiliary Model

- **Assumption:**

  The extended interaction graph $G = (V, E)$ is acyclic.

- **Auxiliary model:** $\Pr\{\tilde{X}_i = +1\} = \Pr\{\tilde{X}_i = -1\} = 1/2$ for $0 \leq i \leq d$.

  ($\tilde{X}_0$: the virtual feature corresponding to the virtual node v_0)

$$\Pr\{\tilde{Y} = +1 | \tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d\} = \sigma\Big(\sum_{i=1}^{d} \beta_i \tilde{X}_0 \tilde{X}_i + \sum_{1 \leq i < j \leq d} \beta_{i,j} \tilde{X}_i \tilde{X}_j\Big)$$

$$\begin{aligned}\Pr\{\tilde{Y} = -1 | \tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d\} &= 1 - \Pr\{\tilde{Y} = +1 | \tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d\} \\ &= \sigma\Big(-\sum_{i=1}^{d} \beta_i \tilde{X}_0 \tilde{X}_i - \sum_{1 \leq i < j \leq d} \beta_{i,j} \tilde{X}_i \tilde{X}_j\Big)\end{aligned}$$

Relationship between Original Model and its Auxiliary Model

- **Original model:**

$$
\begin{aligned}
w_{\{0,i\}} &:= |\Pr(Y = +1|X_i = +1) - \Pr(Y = -1|X_i = +1)| \\
w_{\{i,j\}} &:= |\Pr\left(Y = +1|X_i = +1, X_j = +1\right) \\
&\quad + \Pr\left(Y = +1|X_i = -1, X_j = -1\right) - 1|
\end{aligned}
$$

- **Auxiliary model:**

$$
\tilde{w}_{\{i,j\}} := \left|\Pr(\tilde{\mathrm{Y}} = +1|\tilde{\mathrm{X}}_i = +1, \tilde{\mathrm{X}}_j = +1) - \Pr(\tilde{\mathrm{Y}} = -1|\tilde{\mathrm{X}}_i = +1, \tilde{\mathrm{X}}_j = +1)\right|
$$

Theorem

For $0 \le i < j \le d$,

$$
w_{i,j} = \tilde{w}_{i,j}
$$

Idea of Converting

- Original model and auxiliary model share the same interaction graph.
- Auxiliary model contains only pairwise interactions.
- Assign the empirical weight of the original model into each edge of the auxiliary model.

Detection Algorithm of Extended Interaction Graphs

Algorithm

- *Construct a weighted complete graph* $G' = (V', E')$ *with* $V' = \{v'_0, v'_1, v'_2, \ldots, v'_d\}$.
- *For* $1 \leq i \leq d$, *the weight* $w_{\{0,i\}}$ *of edge* (v'_0, v'_i) *is assigned to be*

$$\left| \frac{4}{n} \sum_{t=1}^{n} \mathbf{1}((\mathrm{x}_i[t], \mathrm{y}[t]) = (+1, +1)) - 1 \right| ;$$

for $1 \leq i < j \leq d$, *the weight* $w_{\{i,j\}}$ *of edge* (v'_i, v'_j) *is assigned to be*

$$\left| \frac{4}{n} \sum_{t=1}^{n} \mathbf{1}((\mathrm{x}_i[t], \mathrm{x}_j[t], \mathrm{y}[t]) = (+1, +1, +1)) + \frac{4}{n} \sum_{t=1}^{n} \mathbf{1}((\mathrm{x}_i[t], \mathrm{x}_j[t], \mathrm{y}[t]) = (-1, -1, +1)) - 1 \right| .$$

Detection Algorithm of Extended Interaction Graphs (Continued)

Algorithm

- *Find a maximum spanning tree* $T' = (V', E_T)$ *of* G' *(by Kruskal's algorithm or Prim's algorithm).*
- *Then the set of the edges in* G *is* $\{(v_i, v_j) : (v'_i, v'_j) \in E_T \text{ and } w_{\{i,j\}} > \gamma'/2\}$, *with*

$$\gamma' = \sqrt{\frac{2}{\pi(d+1)}} \left[\sigma(\lambda + 3\mu) - \sigma(-\lambda + 3\mu)\right].$$

The algorithm is also executed in polynomial time.

Non-Uniform Case

- **Assumption:**
  - $X_1, X_2, \ldots, X_d$ are independent variables with $\Pr\{X_i = +1\} = p_i$, $\Pr\{X_i = -1\} = q_i$ with $p_i + q_i = 1$, for $i = 1, 2, \ldots, d$ (non-uniform features)
  - The interaction graph $G = (V, E)$ is simply a path of length at most 4.
- **Target:**

  Reconstruct the graph from the samples of $(Y, X_1, X_2, \ldots, X_d)$.
- **Construction:**

  Construct a weighted complete graph $G' = (V', E')$ by
  - $V' = (v'_1, v'_2, \ldots, v'_d)$
  - The weight of any edge $(v'_i, v'_j) \in E'$ is assigned to be

$$w_{\{i,j\}} = \Big| Q^{i,j}_{+1,+1,+1} + Q^{i,j}_{-1,-1,+1} + Q^{i,j}_{-1,+1,-1} + Q^{i,j}_{+1,-1,-1} - Q^{i,j}_{+1,+1,-1} - Q^{i,j}_{-1,-1,-1} - Q^{i,j}_{-1,+1,+1} - Q^{i,j}_{+1,-1,+1} \Big|.$$

$(Q^{i,j}_{i_1,i_2,i_3} := \Pr\{Y = i_3 | X_i = i_1, X_j = i_2\})$

Theorem on Detection (Non-uniform Case)

Theorem

Let $T = (V', E_T)$ *be a maximum spanning tree of* G'. *Then*

$$(v_i, v_j) \in E \text{ if and only if } (v_i', v_j') \in E_T \text{ and } w_{\{i,j\}} > 0.$$

Hardness of Detection (Non-uniform Case)

Theorem

Assume that the interaction graph is simply a path of length 5. If the weight of edge (v_i', v_j') in G' is assigned to be

$$w_{\{i,j\}} = \left| \sum_{i_1, i_2, i_3 \in \{+1,-1\}} \alpha_{i_1,i_2,i_3} Q^{i,j}_{i_1,i_2,i_3} \right|,$$

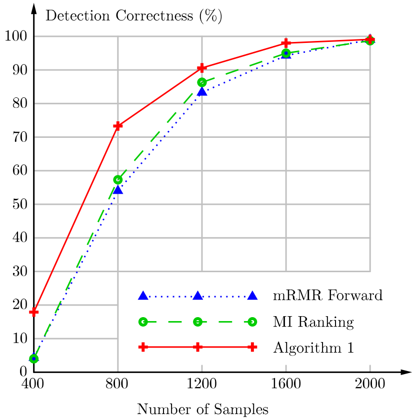
for any constants $\{\alpha_{i_1,i_2,i_3} : i_1, i_2, i_3 \in \{+1,-1\}\}$, then there exists a counterexample where we cannot correctly detect the structure of the interaction graph by finding a maximum spanning tree of G'.

The theorem for the uniform cases cannot be extended into the generic non-uniform cases.

Simulation Experiments

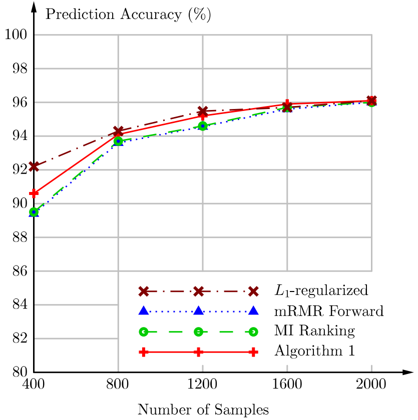
- 1000 logistic regression models
- 15 features, 5 individual effects, 10 pairwise interactions
- 400, 800, 1,200, 1,600, 2,000 samples
- Detection of the interaction graphs

Results of Simulation Experiments - Part 1



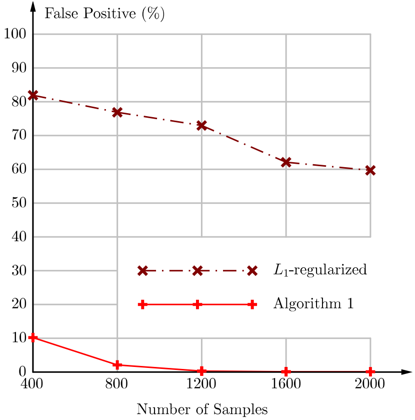
Comparison of detection correctness among mRMR forward selection [Peng, Long & Ding (2005)], feature ranking based on mutual information estimation [Paninski (2003)], and our algorithm.

Results of Simulation Experiments - Part 2



Comparison of prediction correctness among mRMR forward selection [Peng, Long & Ding (2005)], feature ranking based on mutual information estimation [Paninski (2003)], and L_1-penalized logistic regression [Park & Hastie (2007)], and our algorithm.

Results of Simulation Experiments - Part 3



Comparison of false positive rates for detection between L_1-penalized logistic regression [Park & Hastie (2007)] and our Algorithm.

Summary

- Logistic regression models:

$$\Pr\{Y = +1 | X_1, X_2, \ldots, X_d\} = \sigma\Big(\sum_{1 \le i \le d} \beta_i X_i + \sum_{1 \le i < j \le d} \beta_{i,j} X_i X_j\Big)$$

$$\Pr\{Y = -1 | X_1, X_2, \ldots, X_d\} = 1 - \Pr\{Y = +1 | X_1, X_2, \ldots, X_d\}$$

- Interaction graph $G = (V, E)$:

$$(v_i, v_j) \in E \iff \beta_{i,j} \neq 0.$$

- Detection of the interaction graph:
  - Construct a weighted complete graph.
  - Find its maximum spanning tree.
  - Pick the edges with large weights.
- Extended to the models with both individual effects and pairwise interactions

Key References

E. L. Xu, X. Qian, T. Liu, and S. Cui
Detection of Cooperative Interactions in Logistic Regression Models
Submitted to IEEE Transactions on Signal Processing, available at arXiv: 1602.03963

C. K. Chow and C. N. Liu
Approximating Discrete Probability Distributions with Dependence Trees
IEEE Transactions on Information Theory, vol. 14, no. 3, pp. 462-467, May 1968

G. Bresler
Efficiently Learning Ising Models on Arbitrary Graphs
Proceedings in Symposium on Theory of Computing (STOC), Jun. 2015

Key References (Continued)

H. Peng, F. Long, and C. Ding

Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, Aug. 2005

L. Paninski

Estimation of Entropy and Mutual Information

Neural Computation, vol. 15, no. 6, pp. 1191–1253, June 2003

M. Y. Park and T. Hastie

"L_1-regularization Path Algorithm for Generalized Linear Models,"

J. Roy. Stat. Soc. B, vol. 69, no. 4, pp. 659–677, Sept. 2007

Thank you!