Learning over Limited Data Samples: Detection of Cooperative Interactions in Logistic Regression Models

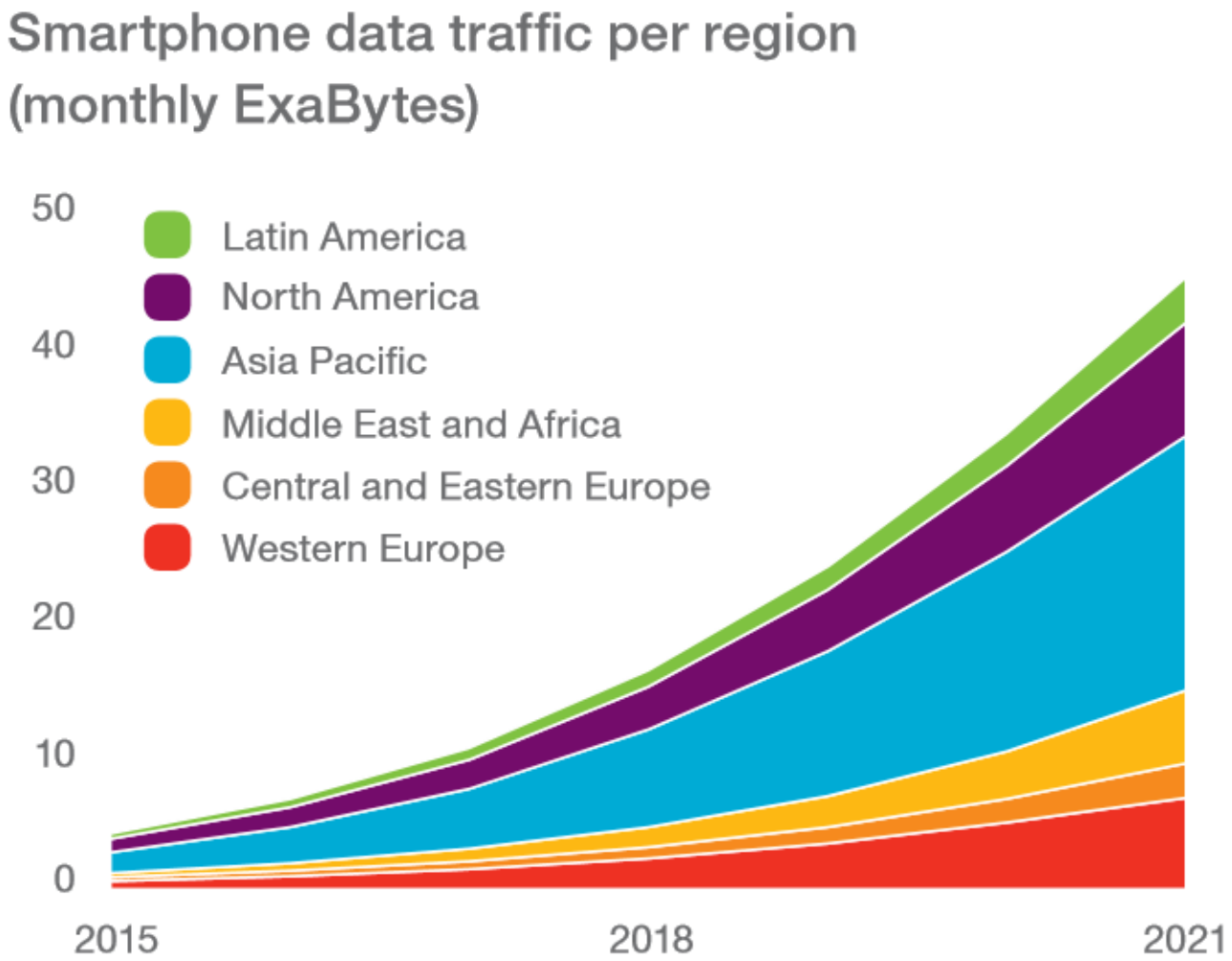
Shuguang (Robert) Cui

University of California-Davis, Professor

(Joint work with Easton Li Xu, Tie Liu, Xiaoning Qian, TAMU)

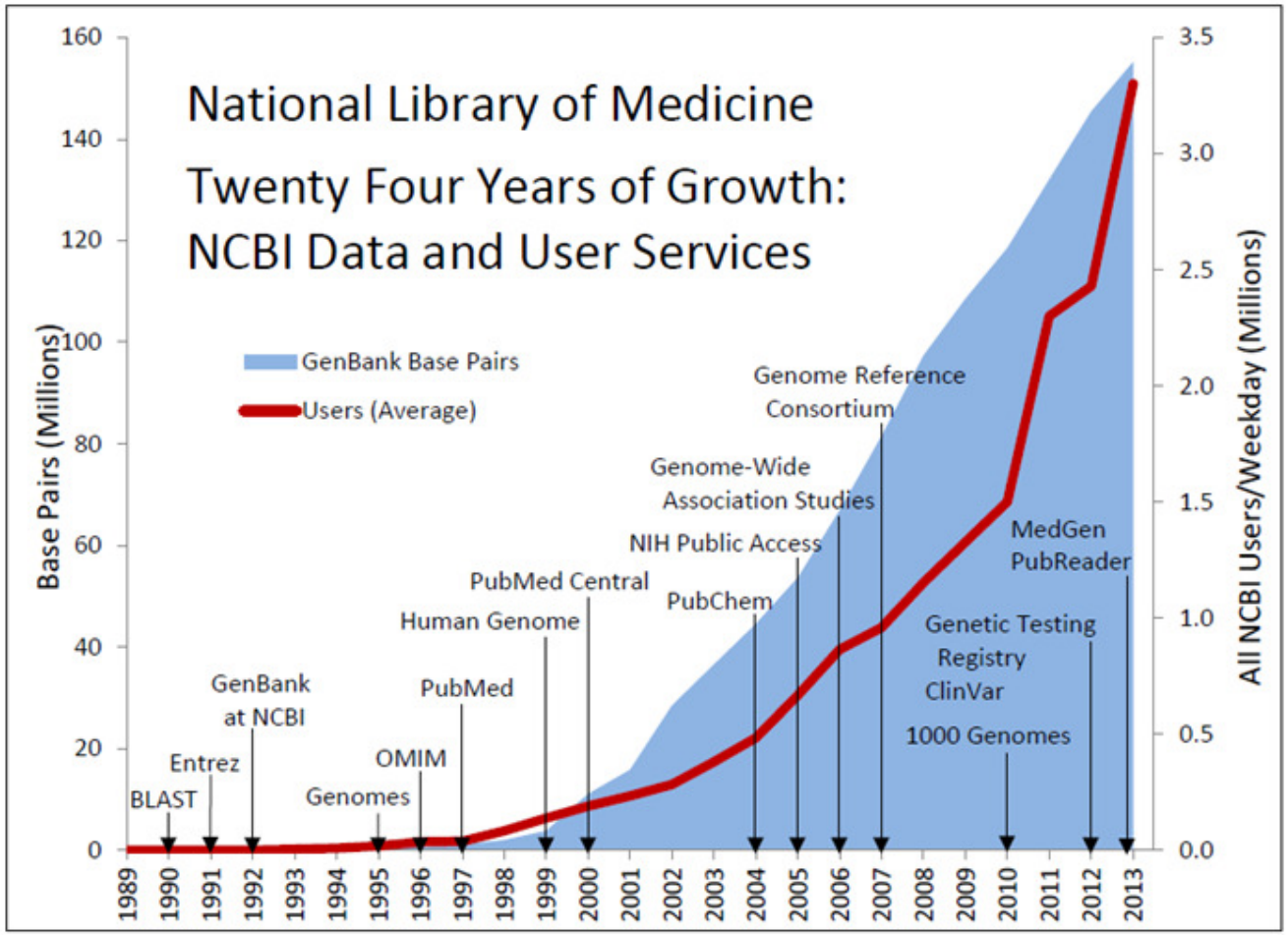
September 23, 2016

Big Data Era: Mobile Data Traffic



Source: http://dazeinfo.com/

Big Data Era: Biomedical Data Growth



Source: https://www.nlm.nih.gov/

Characteristics of Big Data (5V)

- Volume: terabytes (TB), petabytes (PB), exabytes (EB), zettabytes (ZB)...
- Variety: different types like text, audio, image, video...
- Velocity: real-time data
- Veracity: unreliable data quality
- Variability: data inconsistency

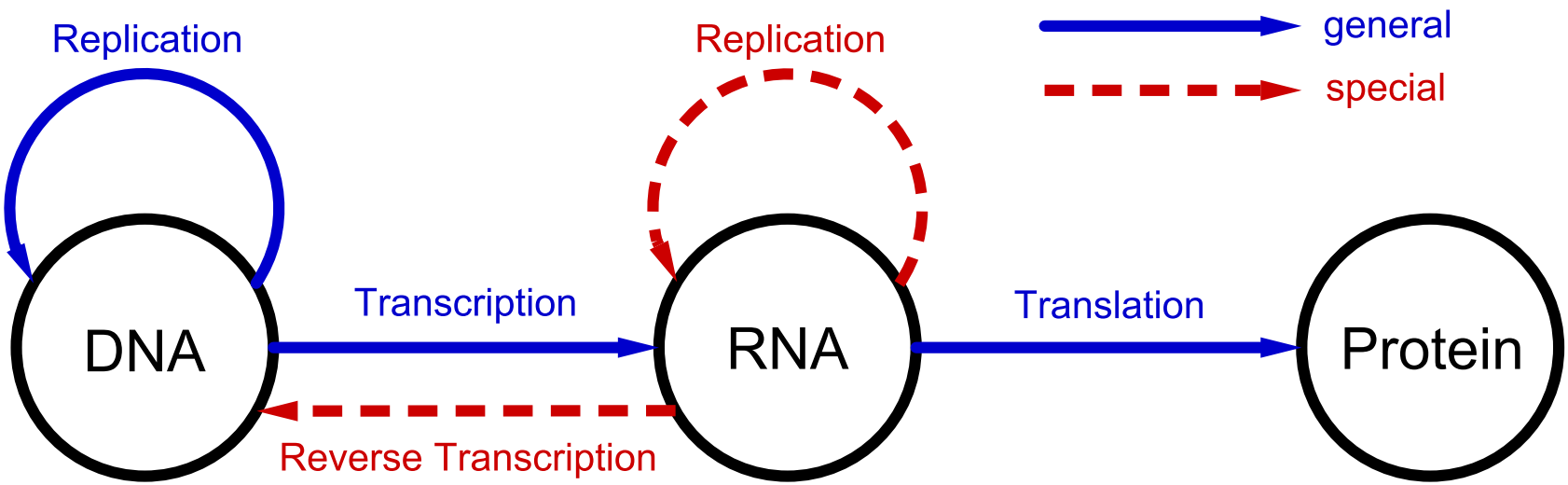
High-Dimensional Data

- Examples:
  - Microarray: Human has 20,000-25,000 genes
  - Mobile user data: more than 0.4 billion users in China Mobile
- Curse of dimensionality:
  - As the dimensionality increases, active data becomes more and more sparse.
  - Processing based on the distance between points becomes less meaningful.
- One solution: feature selection
- New challenge: limited data samples (labeled data)

Outline

- Feature selection, interaction measures, logistic regression
- System models
- Pairwise interaction and interaction graph
- Detection of pairwise interactions
- Detection of pairwise interactions+individual effects

Example in Biology



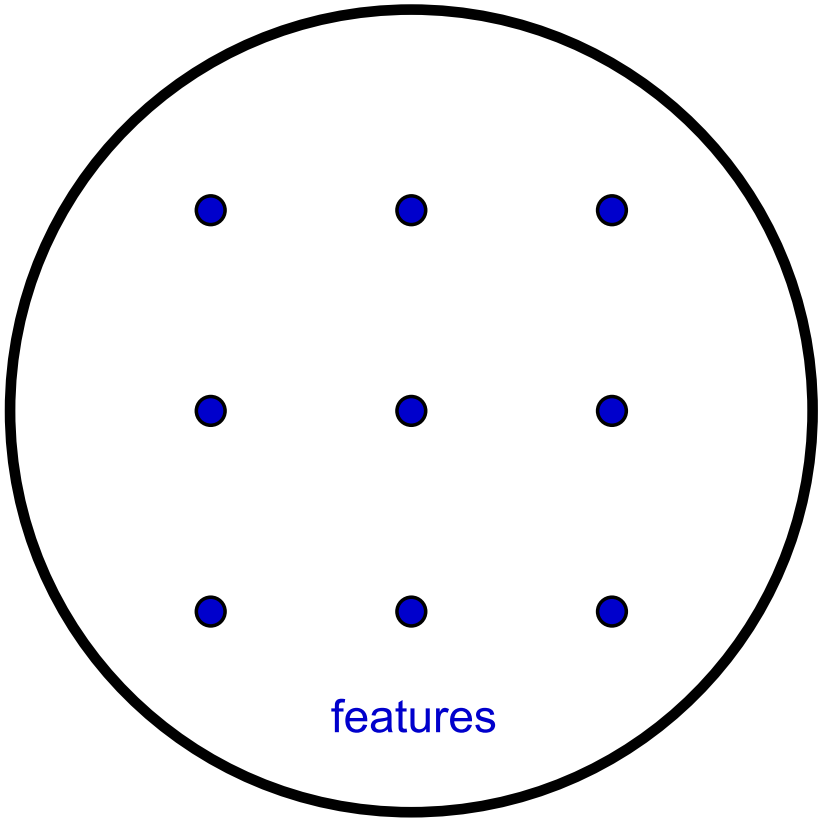
Central Dogma of Molecular Biology

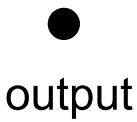
Genes & Diseases

Gene: a region of DNA that serves some function
via encoding the proteins
(more than 20,000 human genes)

Disease: cancer, asthma, cardiopathy...
that are affected by proteins

Feature Selection

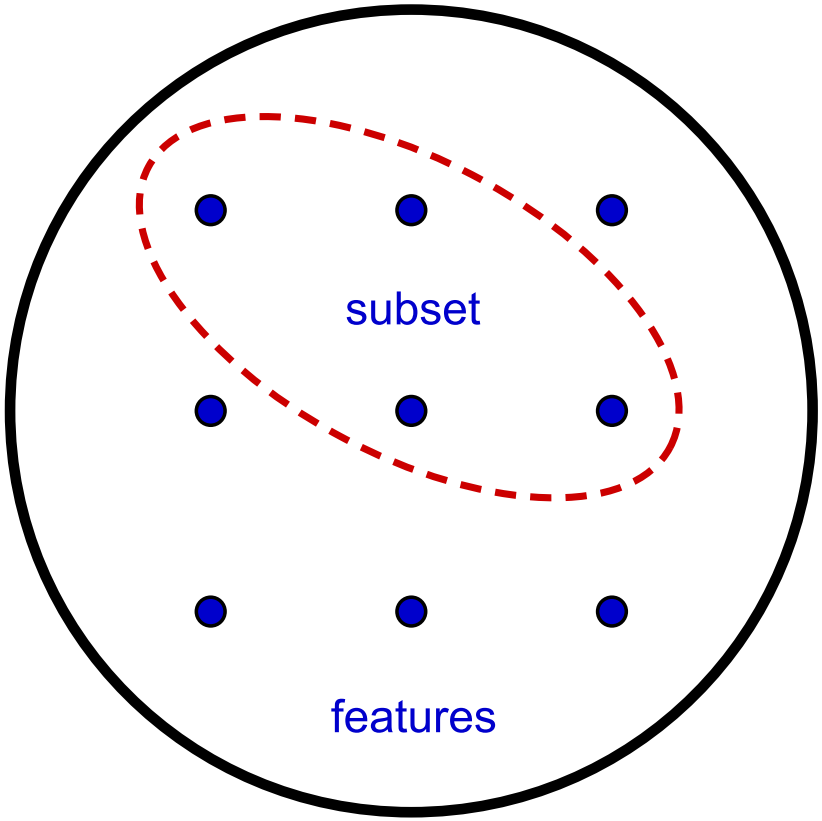


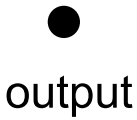


Features: genes,...

Output: healthy or ill, ...

Feature Selection

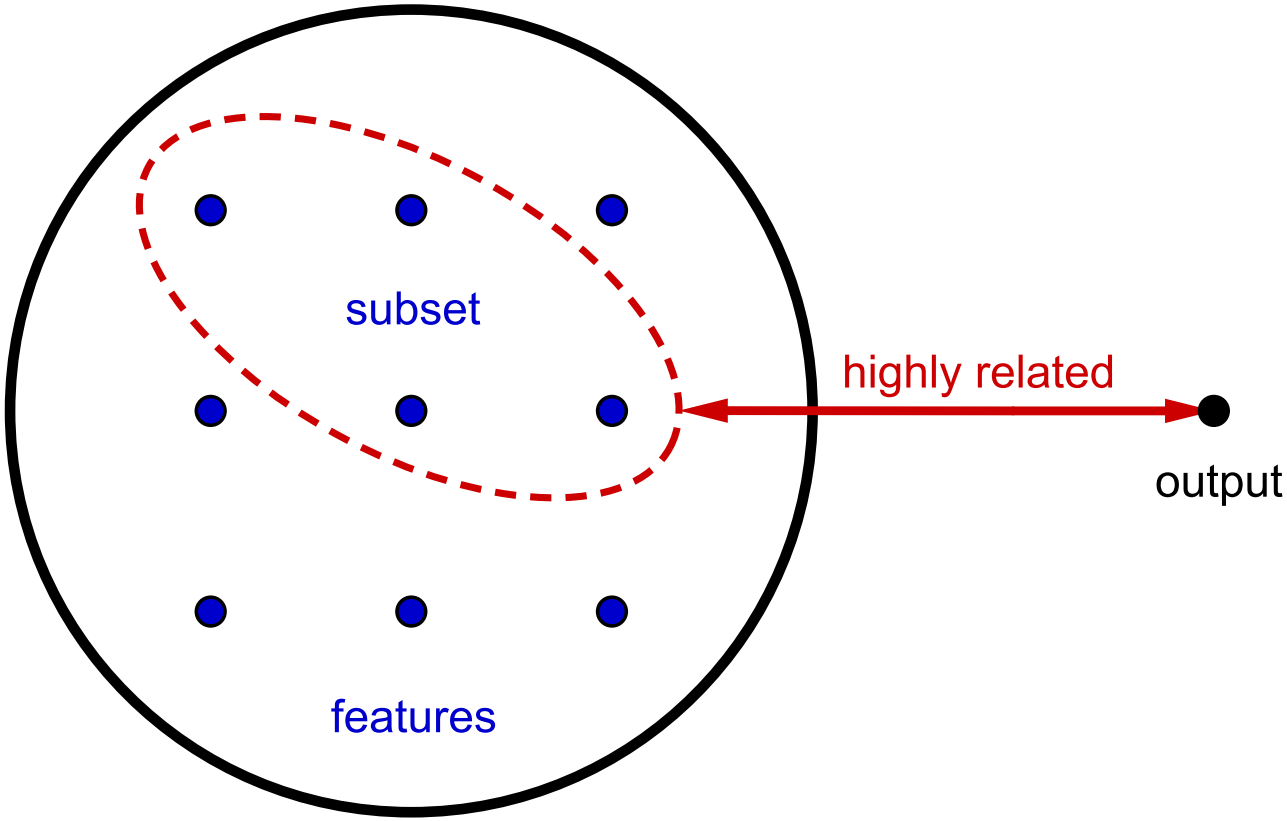




Features: genes,. . .

Output: healthy or ill, . . .

Feature Selection



Features: genes,. . . **Output:** healthy or ill, . . .

Example in Cellular Systems

- System state is user dependent; so each user could be treated as a feature
- Traffic mostly from pairs of nodes
- System states highly dependent on pairwise interactions
- Useful in identifying key pairs of users for various purposes: prioritized services, VIP customer tagging, precision advertisement...

Advantages of Feature Selection

- Simplify model's structure
- Improve model interpretability
- Shorten training time
- Prevent model overfitting

Feature Selection Methods

Three types:

- Filter methods [Ng (1998)]
- Wrapper methods [Kohavi and John (1997)]
- Embedded methods [Breiman, Friedman, Olshen, and Stone (1984)]

Two key questions:

- How to measure relevance?
- How to efficiently (requiring less samples) find the optimal subset?

Individual Feature Selection

1) Pearson's correlation coefficient

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\sigma^2(X)\sigma^2(Y)}}$$

$$\approx \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(x_n - \bar{x})^2} \cdot \sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}}$$

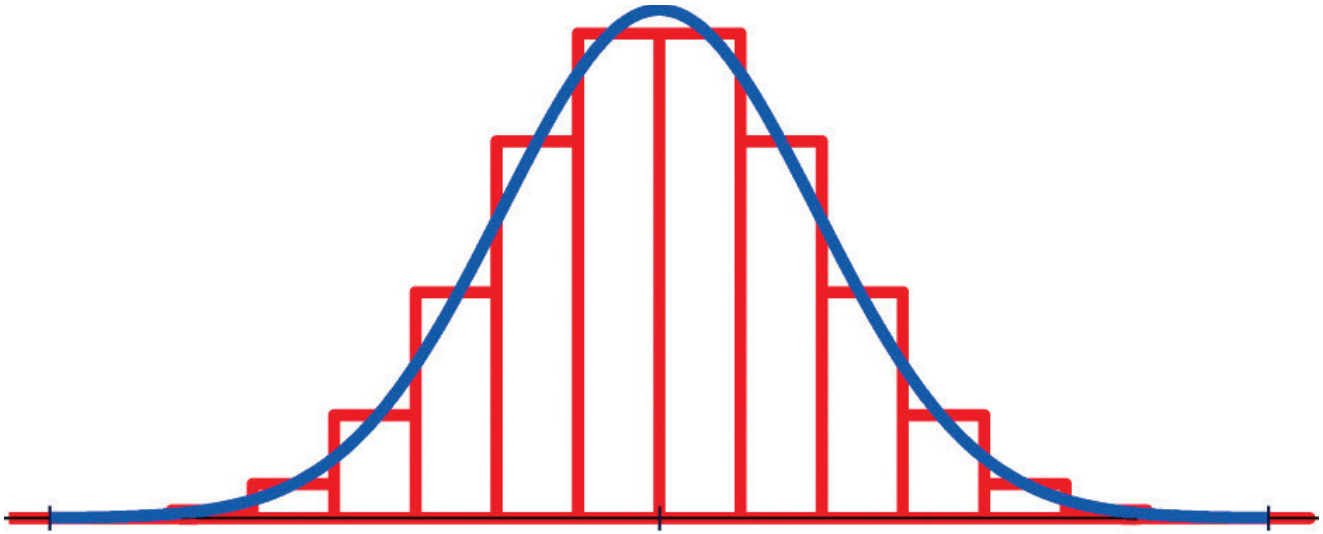
Individual Feature Selection (Continued)

2) Mutual information

$$I(X;Y) = D(p(X,Y)||p(X)p(Y)) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Estimation

$$p(\cdot) \approx \text{ sampling distribution}$$



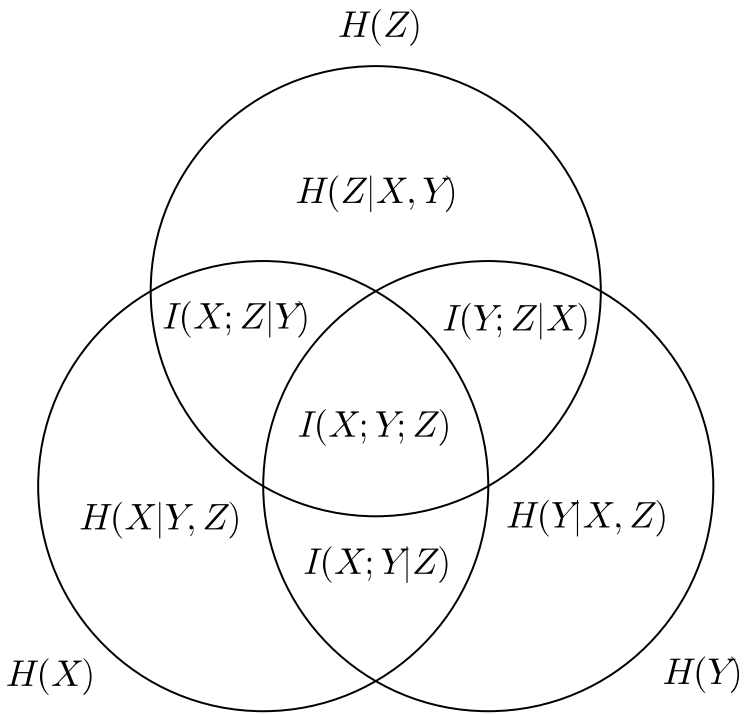
Individual Feature Selection (Continued)

3) Maximal information coefficient (MIC) [Reshef *et. al.* (2011)]

$$MIC(X;Y) = \max_{|\tilde{X}|\cdot|\tilde{Y}|<B} \frac{I(\tilde{X};\tilde{Y})}{\log_2\left[\min\{|\tilde{X}|,|\tilde{Y}|\}\right]}$$

- $\tilde{X}$ and $\tilde{Y}$ discretize X and Y
- $|\tilde{X}|, |\tilde{Y}|$: # of possible values
- $B = N^{0.6}$ (N: # of data samples)
- $0 \leq MIC \leq 1$

Pairwise Feature Interactions



Bivariate synergy [Anastassiou (2007)]

$$SYN(X_1, X_2; Y) = I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y)$$

Our Measure

- Define a new measure called "influence".
- Useful in the detection of the pairwise interactions in logistic regression models.
- Express the pairwise interactions by "interaction graph".
- Solve the case when the interaction graph is acyclic with a polynomial algorithm.
- Require lower order probability distribution information or much less number of samples.

Logistic Regression Models

Linear regression models

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

Logistic regression models

$$\Pr(Y = +1|\mathbf{X}) \sim \boldsymbol{\beta} \cdot \mathbf{X} \text{ and } \Pr(Y = -1|\mathbf{X}) = 1 - \Pr(Y = +1|\mathbf{X})$$

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

$$\Pr(Y = +1|\mathbf{X}) \sim \sigma(\boldsymbol{\beta} \cdot \mathbf{X}) \text{ and } \Pr(Y = -1|\mathbf{X}) = 1 - \Pr(Y = +1|\mathbf{X})$$

Pairwise Interaction

Logistic regression model with individual effects and pairwise interactions

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

$$\Pr(Y = -1|\mathbf{X}) = 1 - \Pr(Y = +1|\mathbf{X})$$

System Model

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

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

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

  - $\sigma(x) := 1/(1 + e^{-x})$: the sigmoid function
  - β_i: the parameter corresponding to the individual effect X_i for $1 \le i \le d$.
  - $\beta_{i,j}$: the parameter corresponding to the covariate pair $\{X_i, X_j\}$ for $1 \le i < j \le d$.

Individual Effects and Pairwise Interactions

- **Definitions:**
  - $\beta_i \neq 0$: 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.

- **Target:**

  Detect all individual effects and pairwise interactions in logistic regression models.

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

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

$$f(x_1, x_2, \ldots, x_p) = \exp\left\{\sum_{i \in V} \beta_i x_i + \sum_{\{i,j\} \in E} \beta_{i,j} x_i x_j - \Phi(\beta)\right\}$$

- 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
- $\alpha \leq |\beta_{i,j}| \leq \beta$, $|\beta_i| \leq h$.

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

Theorem (Bresler 2015)

Let $\delta = \frac{1}{2}e^{-2(\beta p+h)}$, $\tau^* = \frac{\alpha^2\delta^{4p+1}}{16p\beta}$, $\epsilon^* = \frac{\tau^*}{2}$, $\ell^* = \frac{8}{(\tau^*)^2}$. *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:

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

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)

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 the maximum spanning tree T of G (by Kruskal's algorithm or Prim's algorithm).
- Set a node v to be the root of T, then rank the other nodes by their depths.

Our Work

- Model all pairwise interactions by a so-called interaction graph; treat individual effect as special cases.
- 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.
- Use "influence" as the measure of correlation between the pairs and the outcome, requiring lower-oder probabilities only.
- 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 features 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 $\beta_{i,j} = 0$

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

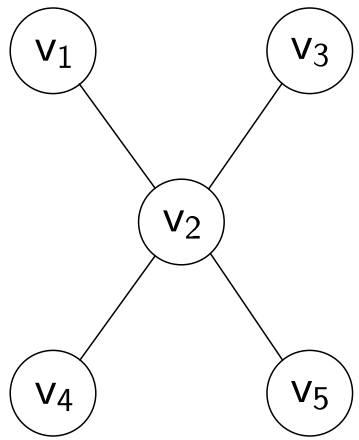
Interaction Graph

Interaction Graph: Let $G = (V, E)$ be the interaction graph with $V = \{v_1, v_2, \ldots, v_d\}$, 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:

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

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



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

Assumption, Difficulty & Target

- **Assumption:**

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

  - When there are 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.

- **Difficulty:**

  We don't know which edges this graph has.

- **Target:**

  Detect the structure of the interaction graph from limited samples.

Construction of a Weighted Complete Graph

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

$$w_{\{i,j\}} = f\left(\left|Q^{i,j}_{1,1,1} - \frac{1}{2}\right|\right).$$

Here,

- $Q^{i,j}_{i_1,i_2,i_3} \triangleq \Pr(Y = i_3 | X_i = i_1, X_j = i_2)$
- f: a nonnegative, strictly increasing function on $[0, 1/2]$.

e.g. (1) $f(x) \triangleq 2x$:

$$w_{\{i,j\}} = \left|Q^{i,j}_{1,1,1} - Q^{i,j}_{1,1,-1}\right|$$

(2) $f(x) \triangleq (1/2 + x)\log(1 + 2x) + (1/2 - x)\log(1 - 2x)$ and $0 \log 0 \triangleq 0$:

$$w_{\{i,j\}} = 3\log 2 + \sum_{i_1,i_2,i_3 \in \{1,-1\}} \frac{Q^{i,j}_{i_1,i_2,i_3}}{4} \log \frac{Q^{i,j}_{i_1,i_2,i_3}}{4}.$$

Best Choice of the Weights

We choose

$$w_{\{i,j\}} = \left| Q^{i,j}_{1,1,1} - Q^{i,j}_{1,1,-1} \right| \text{ (influence).}$$

- Easily estimated by the empirical distributions of n samples $(X_1[t], X_2[t], \ldots, X_d[t], Y[t])$ for $1 \leq t \leq n$:

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

- It has a lower estimation error than other measures with more complicated forms.

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

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

Theorem on Detection (Case 1)

Theorem

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

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

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 the maximum spanning tree* $T' = (V', E_T)$ *of* G' *(by Kruskal's algorithm or Prim's algorithm).*
- *Then the edges in* G *are* $\{(v_i, v_j) : (v'_i, v'_j) \in E_T$ *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 $(X_1[t], X_2[t], \ldots, X_d[t], Y[t])$ for $1 \leq t \leq n$, we estimate

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

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

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*

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

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: $\Theta\left(d^3 \log \frac{d}{\epsilon}\right)$

Models with both Individual Effects and Pairwise Interactions

- **For example:**
  - 4 features 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 $G = (V, E)$,

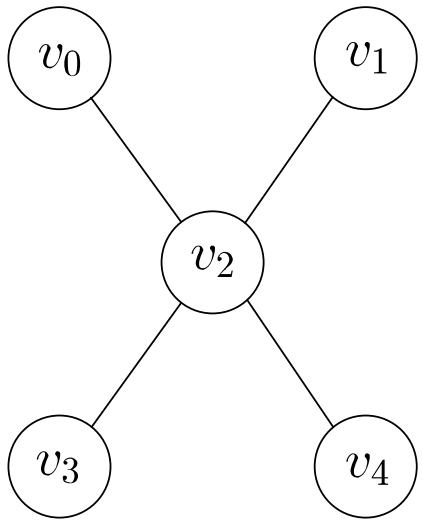
- $V = \{v_0(\text{virtual vertex}), v_1, v_2, \ldots, v_d\}$
- $(v_0, v_i) \in E$ if and only if X_i has an individual effect
- $(v_i, v_j) \in E$ if and only if X_i and X_j have a cooperative interaction

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

For example:

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

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



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

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)

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

$$\Pr\{\tilde{Y} = -1 | \tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d\} = 1 - \Pr\{\tilde{Y} = +1 | \tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d\}$$

$$= \sigma\Big(-\sum_{i=1}^{d} \beta_i \tilde{X}_0 \tilde{X}_i - \sum_{1 \leq i < j \leq d} \beta_{i,j} \tilde{X}_i \tilde{X}_j\Big)$$

Relationship between Original Model and its Auxiliary Model

- **Original model:**

$$w_{\{0,i\}} := |4\Pr(X_i = +1, Y = +1) - 1|$$
$$w_{\{i,j\}} := |4\Pr(X_i = +1, X_j = +1, Y = +1) + 4\Pr(X_i = -1, X_j = -1, Y = +1) - 1|$$

- **Auxiliary model:**

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

Theorem

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

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

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*

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

Detection Algorithm of Extended Interaction Graphs (Continued)

Algorithm

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

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

The algorithm is also executed in polynomial time.

Advantages of Our Measure

- **Estimation aspect:**
  It is easier to estimate and analyze a measure only related to lower-order conditional probabilities than other entropy-related measures.

- **Detection aspect:**
  It can be used to detect all pairwise interactions with a low error probability, better than other known algorithms with other measures.

Non-Uniform Case

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

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

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

Theorem on Detection (Non-uniform Case)

Theorem

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

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

When the interaction graph $G = (V, E)$ is simply a path of length at most 4,

edges in the interaction graph

$$\Updownarrow$$

non-zero weighted edges in the maximum spanning tree

Hardness on Detection (Non-uniform Case)

Theorem

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

$$w_{\{i,j\}} = \max_{i_1,i_2,i_3 \in \{+1,-1\}} Q^{i,j}_{i_1,i_2,i_3} - \min_{i_1,i_2,i_3 \in \{+1,-1\}} Q^{i,j}_{i_1,i_2,i_3}.$$

or

$$\begin{aligned} w(v_i', v_j') = \Big| & Q^{i,j}_{+1,+1,+1} + Q^{i,j}_{-1,-1,+1} + Q^{i,j}_{-1,+1,-1} + Q^{i,j}_{+1,-1,-1} \\ & - Q^{i,j}_{+1,+1,-1} - Q^{i,j}_{-1,-1,-1} - Q^{i,j}_{-1,+1,+1} - Q^{i,j}_{+1,-1,+1} \Big|, \end{aligned}$$

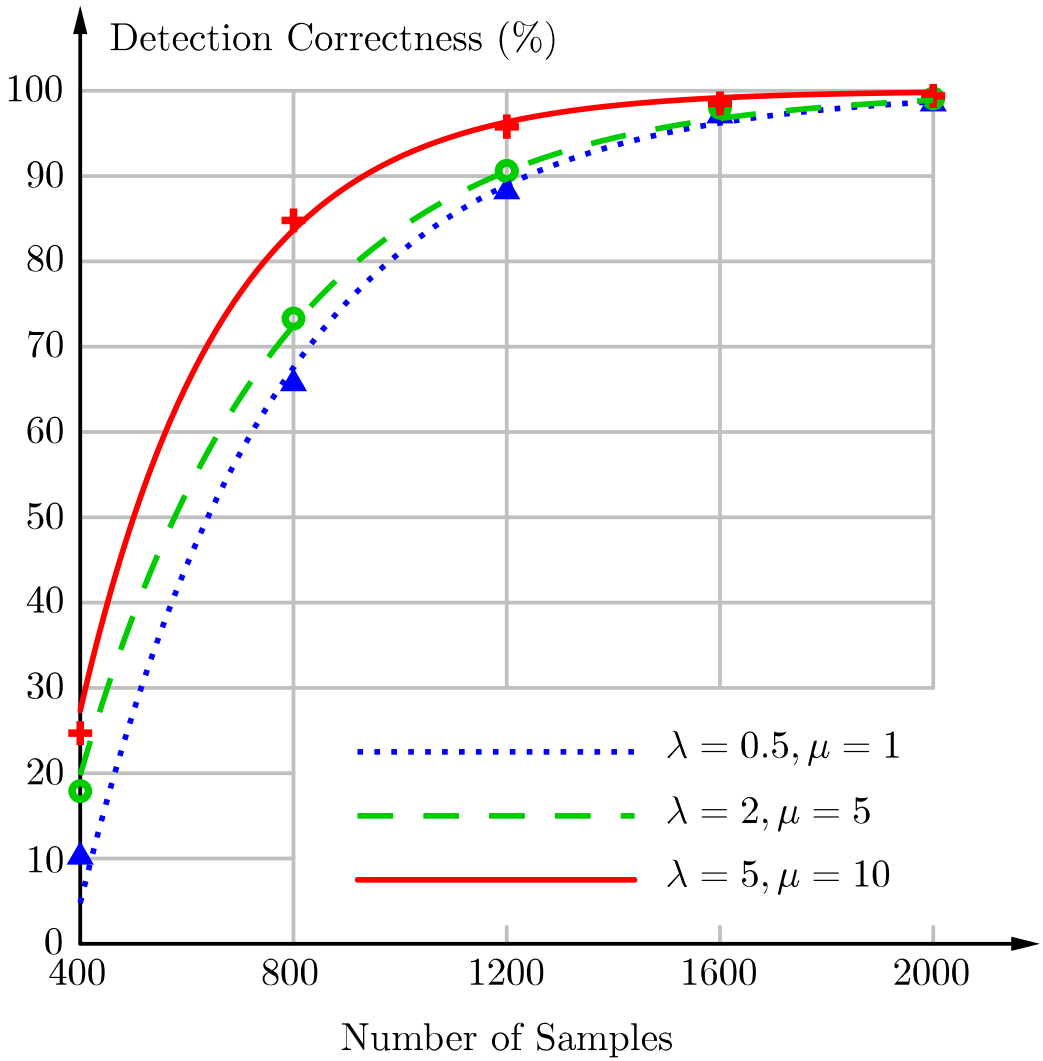
then there exists a counterexample where we cannot correctly detect the structure of the interaction graph by finding the maximum spanning tree of G'.

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

Simulation Experiments

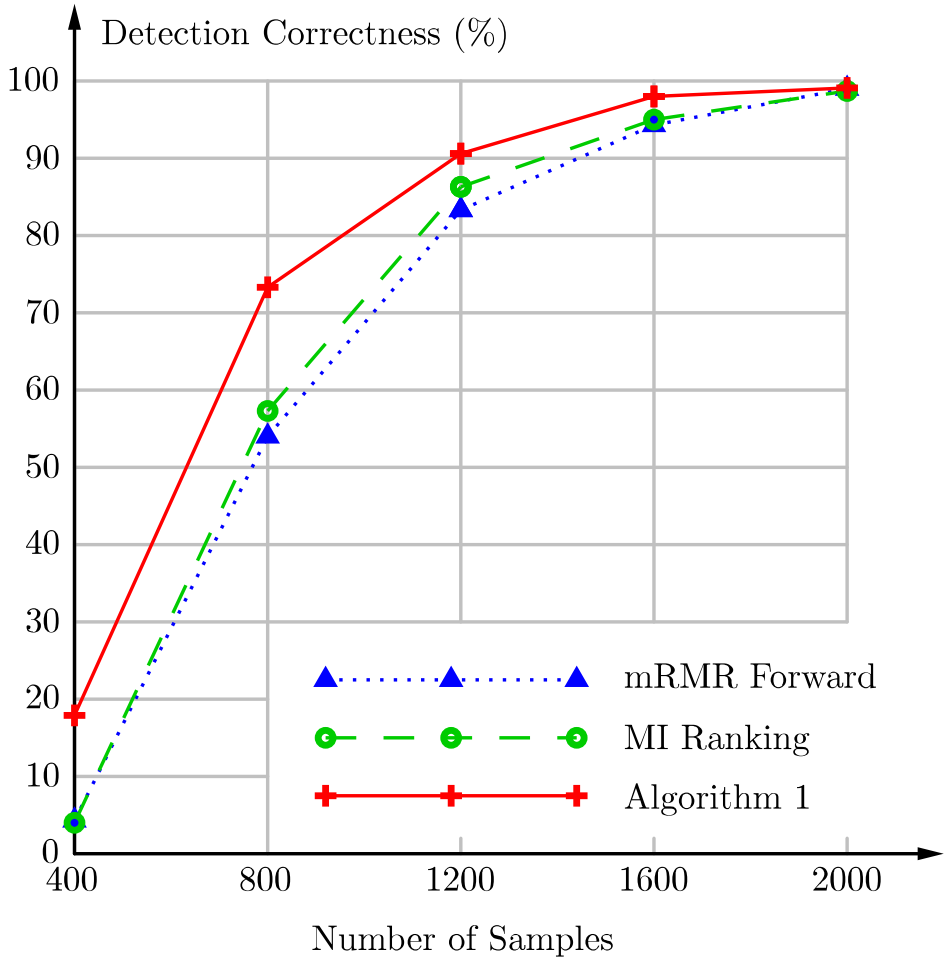
- 1000 logistic regression models
- 15 features, 5 individual effects, 10 pairwise interactions
- 400, 800, 1200, 1600, 2000 samples
- Detection of the interaction graphs

Results of Simulation Experiments - Part 1



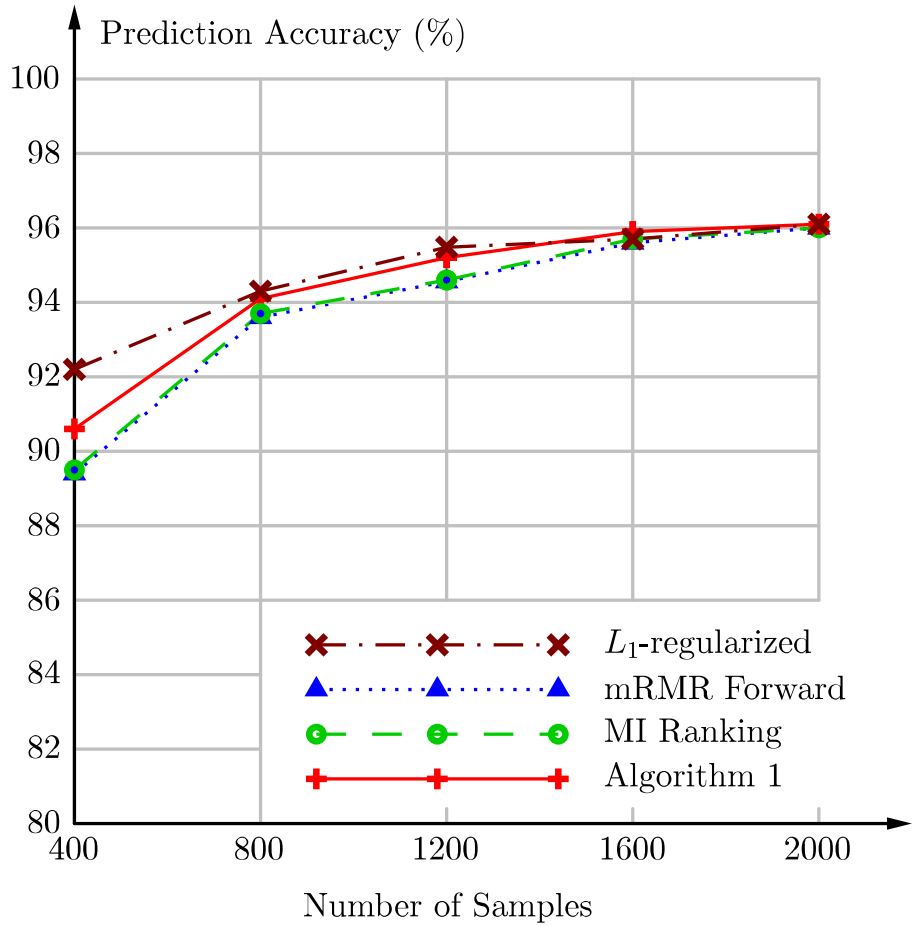
Detection correctness by our algorithm using different parameter ranges $[-\mu, -\lambda] \cup [\lambda, \mu]$.

Results of Simulation Experiments - Part 2



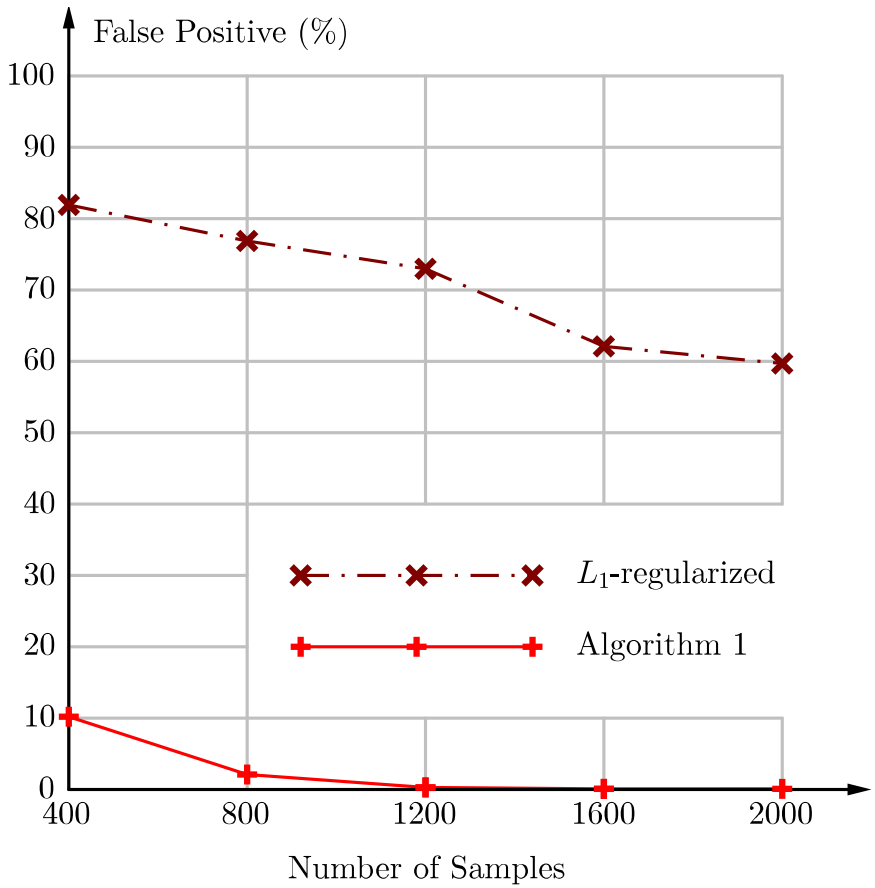
Comparison of detection correctness among mRMR forwarding selection [Peng, Long & Ding (2005)], feature ranking based on mutual information estimation [Paninski (2003)], and our algorithm.

Results of Simulation Experiments - Part 3



Comparison of prediction correctness among mRMR forwarding 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 4



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

Conclusion

- 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 $G = (V, E)$:

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

- Pairwise interaction measure: influence
- Detection of the interaction graph:
  - Construct a weighted 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

D. N. Reshef, Y. Reshef, H. K. Finucane, S. R. Grossman, and G. McVean

Detecting Novel Associations in Large Data Sets

Science, vol. 334, no. 8, pp. 1518-1524, Dec. 2011

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, 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.

D. Anastassiou

Computational Analysis of the Synergy among Multiple Interacting Genes

Molecular Systems Biology, vol. 3, no. 83, 2007

Key References (Continued)

A. Y. Ng,

On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples

Proc. 15th Int. Conf. Mach. Learn. pp. 404-412, San Francisco, CA, 1998.

R. Kohavi and G. John

Wrappers for Feature Selection

Artif. Intell., vol. 97, no. 1-2, pp. 273-324, 1997.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone

Classification and Regression Trees

CRC Press, 1984.

About My Group

- 10 Ph.D students and 2 postdocs;
- 6 DoD and 9 NSF federal grants so far;
- Current research interests:
  - Large scale data analytics;
  - Big data aware next-generation wireless;
  - Cognitive wireless communication networks;
  - Convex optimization;
  - Design synergy of system protocols and hardware;

Thank you!