Sliced Inverse Regression with Interaction (Siri) Detection for non-Gaussian BN learning

Jun S. Liu
Department of Statistics
Harvard University

Joint work with Bo Jiang
## General: Regression and Classification

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind 1</td>
<td>$x_{11}, x_{12}, \ldots, x_{1p}$</td>
</tr>
<tr>
<td>Ind 2</td>
<td>$x_{21}, x_{22}, \ldots, x_{2p}$</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>Ind N</td>
<td>$x_{N1}, x_{N2}, \ldots, x_{NP}$</td>
</tr>
</tbody>
</table>
THE LAB WHERE
THEY STUDY DRUG INTERACTION
Variable Selection with Interaction

Let $Y \in \mathbb{R}$ be a univariate response variable and $X \in \mathbb{R}^p$ be a vector of $p$ continuous predictor variables

$$Y = X_1 \times X_2 + \epsilon, \; \epsilon \sim N(0, \sigma^2), \; X \sim \text{MVN}(0, I_p)$$

Suppose $p = 1000$. How to find $X_1$ and $X_2$?

One step forward selection $\sim 500,000$ interaction terms
Let $Y \in R$ be a univariate response variable and $X \in R^p$ be a vector of $p$ continuous predictor variables.

Suppose $p = 1000$. How to find $X_1$ and $X_2$?

One step forward selection:\ $\sim 500,000$ interaction terms

Is there any marginal relationship between $Y$ and $X_1$?
[Y|X] ? [X|Y] ? Who is behind the bar?

How long should I wait before telling him that he’s on the outside of the cage, not the inside?
## General: Regression and Classification

### Covariates

<table>
<thead>
<tr>
<th>Ind 1</th>
<th>( x_{11}, x_{12}, \ldots, x_{1p} )</th>
<th>( Y_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind 2</td>
<td>( x_{21}, x_{22}, \ldots, x_{2p} )</td>
<td>( Y_2 )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Ind N</td>
<td>( x_{N1}, x_{N2}, \ldots, x_{NP} )</td>
<td>( Y_N )</td>
</tr>
</tbody>
</table>

### Responses

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}
\]

**How to model this?**
Naïve Bayes model

\[
P(Y|X_1, \ldots, X_m) = \frac{P(Y) \prod_{j=1}^{m} P(X_j|Y)}{P(X_1, \ldots, X_m)}.
\]
(Augmented) Naïve Bayes Model

- (Augmented) Naïve Bayes Classifier with Variable Selection and Interaction Detection (Yuan Yuan et al.): discrete univariate response and continuous (but discretized) predictors
- Bayesian Partition Model for eQTL study (Zhang et al. 2010): continuous multivariate responses and discrete predictors
- Sliced Inverse Regression with Interaction Detection (SIRI): continuous univariate response and continuous predictors
Tree-Augmented Naïve Bayes

(TAN) (tree-augmented naïve Bayes)
Augmented Naïve Bayes

\[ Y \]

\[ X_{01}, X_{02} \]

\[ X_{11}, X_{12}, X_{13} \]

\[ X_{2.11}, X_{2.12}, X_{2.13} \]

\[ X_{2.21}, X_{2.22} \]
How about continuous covariates?

- We may discretize $Y$, and discretize each $X$
- Or discretize $Y$, assuming joint Gaussian distributions on $X$?
- Sound familiar?
An observation:

\[ Y = X_1 \times X_2 + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad X \sim \text{MVN}(0, I_p) \]
Sliced Inverse Regression (SIR, Li 1991)

SIR is a tool for dimension reduction in multivariate statistics. Let $Y \in \mathbb{R}$ be a univariate response variable and $X \in \mathbb{R}^p$ be a vector of $p$ continuous predictor variables.

$$Y = f(\theta_1^T X, ..., \theta_K^T X, \epsilon)$$

$f$ is an unknown function and $\epsilon$ is the error with finite variance. How to identify unknown projection vectors $\theta_1, ..., \theta_K$?
SIR Algorithm

Let $\Sigma_{xx}$ be the covariance matrix of $X$. Standardize $X$ to:

$$ Z = \Sigma_{xx}^{-1/2} \{ X - E(X) \} $$

Divide the range of $y_i$ into $S$ nonoverlapping slices $H_s \in \{1, \ldots, S\}$

$n_s$ is the number of observations within each slice

Compute the mean of $z_i$ over all slices $z_s = n_s^{-1} \sum_{i \in H_s} z_i$, and calculate the estimate for $Cov\{E(X|Y)\}$:

$$ \hat{M} = n^{-1} \sum_{s=1}^{S} n_s^{-1} z_s z_s^T $$

Identify largest $K$ eigenvalues of $\hat{M}$, $\hat{\lambda}_k$ and corresponding eigenvectors $\hat{\eta}_k$. Then,

$$ \hat{\theta}_k = \Sigma_{xx}^{-1/2} \hat{\eta}_k \quad (k = 1, \ldots, K) $$
SIR with Variable Selection

Only a subset of predictors are relevant: $\beta_1, \ldots, \beta_K$ are sparse

Backward subset selection (Cook 2004, Li et al. 2005)

Shrinkage estimates of $\beta_1, \ldots, \beta_K$ using $L_1$- or $L_2$-penalty:

Regularized SIR (RSIR, Zhong et al. 2005)

Sparse SIR (SSIR, Li 2007)

Correlation Pursuit (Zhong et al. 2012) : A forward selection and backward elimination procedure motivated by F-test in stepwise regression

$$F_{1, n- d- 1} = (n- d- 1) \frac{\hat{R}_{d+ 1}^2 - \hat{R}_d^2}{1 - \hat{R}_{d+ 1}^2}$$
Correlation Pursuit (COP)

Let $A$ be the current set of selected predictors and $\hat{\lambda}_k^A$ the $k$th largest eigenvalue estimated by SIR based on predictors in $A$. For $j$th predictor ($j \notin A$), $X_j$, define statistic

$$COP_{k}^{A+ j} = n \frac{(\lambda_k^{A+ j} - \hat{\lambda}_k^A)}{1 - \hat{\lambda}_k^{A+ j}}$$
Correlation Pursuit (COP)

Let $A$ be the current set of selected predictors and $\hat{\lambda}_k^A$ the $k$th largest eigenvalue estimated by SIR based on predictors in $A$. For $j$th predictor ($j \notin A$), $X_j$, define statistic

$$COP_{k}^{A+ j} = n \frac{\left(\lambda_k^{A+ j} - \hat{\lambda}_k^A\right)}{1 - \hat{\lambda}_k^{A+ j}}$$

If $j \notin A$, $COP_{k}^{A+ j}$ ($k = 1, \ldots, K$) are asymptotically i.i.d. $\chi^2(1)$, and $COP_{1:K}^{A+ j} = \sum_{k=1}^{K} COP_{k}^{A+ j}$ is asymptotically $\chi^2(K)$.
Correlation Pursuit (COP)

Let $A$ be the current set of selected predictors and $\hat{\lambda}_k^A$ the $k$th largest eigenvalue estimated by SIR based on predictors in $A$.

For $j$th predictor ($j \notin A$), $X_j$, define statistic

$$
COP_{k}^{A^+j} = n \frac{(\lambda_k^{A^+j} - \hat{\lambda}_k^A)}{1 - \hat{\lambda}_k^{A^+j}}
$$

If $j \notin A$, $COP_{k}^{A^+j} (k = 1, \ldots, K)$ are asymptotically i.i.d. $\chi^2(1)$, and

$$
COP_{1:K}^{A^+j} = \sum_{k=1}^{K} COP_{k}^{A^+j}
$$
is asymptotically $\chi^2(K)$.

The stepwise procedure is consistent if $p = O(n^r), r < 1/2$. 
Correlation Pursuit (COP)

Let $A$ be the current set of selected predictors and $\hat{\lambda}_k^A$ the $k$ th largest eigenvalue estimated by SIR based on predictors in $A$

For $j$ th predictor ($j \notin A$), $X_j$, define statistic

$$COP_{k}^{A^+ j} = n \frac{(\hat{\lambda}_k^{A^+ j} - \hat{\lambda}_k^A)}{1 - \hat{\lambda}_k^{A^+ j}}$$

If $j \notin A$, $COP_{k}^{A^+ j} (k = 1, \ldots, K)$ are asymptotically i.i.d. $\chi^2(1)$, and $COP_{1:K}^{A^+ j}$ is asymptotically $\chi^2(K)$

The stepwise procedure is consistent if $p = O(n^{r})$, $r < 1/2$

Dimension $K$ and threshold in forward selection (backward elimination) are chosen by cross-validation
SIR via MLE

Let $A$ be the set of relevant predictors and $C = A^C$, $d = |A|$

$X_A | Y \in H_s \sim N(\mu_s, \Sigma)$
SIR via MLE

Let $A$ be the set of relevant predictors and $C = \neg A$, $d = |A|$

\[ X_A | Y \in H_s \sim N(\mu_s, \Sigma) \]

\[ X_c | X_A, Y \in H_s \sim N(X_A \theta, \Sigma_0) \]
SIR via MLE

Let $A$ be the set of relevant predictors and $C = A^C$, $d = |A|$

$$X_A|Y \in H_s \sim N(\mu_s, \Sigma)$$

$$X_c|X_A, Y \in H_s \sim N(X_A \theta, \Sigma_0)$$

$$\mu_s = \alpha + \Gamma \gamma_s$$, where $\gamma_s \in R^K$ and $\Gamma$ is a $d \times K$ orthogonal matrix
SIR via MLE

Let $A$ be the set of relevant predictors and $C = A^C$, $d = |A|

\begin{align*}
X_A | Y \in H_s & \sim N(\mu_s, \Sigma) \\
X_c | X_A, Y \in H_s & \sim N(X_A \theta, \Sigma_0)
\end{align*}

$\mu_s = \alpha + V^K$, belongs to a $K$-dimensional affine space ($K < d$)
SIR via MLE

Let $A$ be the set of relevant predictors and $C = A^C$, $d = |A|$

$$X_A | Y \in H_s \sim N(\mu_s, \Sigma)$$

$$X_C | X_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + \mathcal{V}^K$, belongs to a $K$-dimensional affine space ($K < d$)

MLE of the span of subspace $\mathcal{V}^K$ coincides with SIR directions

(Cook 2007, Szretter and Yohai 2009)
SIR via MLE

Let $A$ be the set of relevant predictors and $C = \neg A$, $d = |A|$

$$X_A|Y \in H_s \sim N(\mu_s, \Sigma)$$

$$X_c|X_A, Y \in H_s \sim N(X_A \theta, \Sigma_0)$$

$\mu_s = \alpha + V^K$, belongs to a $K$-dimensional affine space ($K < d$)

MLE of the span of subspace $V^K$ coincides with SIR directions (Cook 2007, Szretter and Yohai 2009)

Given current $A$ and predictor $X_j \notin A$, we want to test $H_0: X_j$ is irrelevant, vs. $H_1: X_j$ is relevant
SIR via MLE

Let $A$ be the set of relevant predictors and $C = A^C$, $d = |A|$  

$$X_A|Y \in H_s \sim N(\mu_s, \Sigma)$$

$$X_c|X_A, Y \in H_s \sim N(X_A \beta, \Sigma_0)$$

$\mu_s = \alpha + V^K$, belongs to a $K$-dimensional affine space ($K < d$)

MLE of the span of subspace $V^K$ coincides with SIR directions (Cook 2007, Szretter and Yohai 2009)

Given current $A$ and predictor $X_{j \notin A}$, we want to test

$H_0 : X_j$ is irrelevant, vs. $H_1 : X_j$ is relevant

$$\frac{P_{M_1}(X|Y)}{P_{M_0}(X|Y)} = \frac{P_{M_1}(X_j|X_A, Y)}{P_{M_0}(X_j|X_A, Y)}$$
SIR via MLE

Let $A$ be the set of relevant predictors and $C = A^C$, $d = |A|$.

$X_A | Y \in H_s \sim N(\mu_s, \Sigma)$

$X_c | X_A, Y \in H_s \sim N(X_A \theta, \Sigma_0)$

$\mu_s = \alpha + V^K$, belongs to a $K$-dimensional affine space ($K < d$)

MLE of the span of subspace $V^K$ coincides with SIR directions (Cook 2007, Szretter and Yohai 2009)

Given current $A$ and predictor $X_{j \notin A}$, we want to test

$H_0 : X_j$ is irrelevant, vs. $H_1 : X_j$ is relevant

$$LR_j = \frac{P_{\hat{M}_1}(X_j|X_A, Y)}{P_{\hat{M}_0}(X_j|X_A, Y)}$$
LR Test vs. COP

Given current $A$, the likelihood ratio (LR) test statistic of $H_0 : X_j$ is irrelevant, vs. $H_1 : X_j$ is relevant

$$2 \text{LR}_j = - n \left( \sum_{k=1}^{K} \log \left( 1 - \hat{\lambda}_{k A_j} \right) - \sum_{k=1}^{K} \log \left( 1 - \hat{\lambda}_k \right) \right)$$

$$= n \sum_{k=1}^{K} \log \left( 1 + \frac{\hat{\lambda}_{k A_j} - \hat{\lambda}_{A_k}}{1 - \hat{\lambda}_{k A_j}} \right)$$

Under $H_0 : X_j$ is irrelevant

$$\text{COP}_{k A_j} = n \frac{\left( \hat{\lambda}_{k A_j} - \hat{\lambda}_{A_k} \right)}{1 - \hat{\lambda}_{k A_j}} \frac{\left( \hat{\lambda}_{k A_j} - \hat{\lambda}_{A_k} \right)}{1 - \hat{\lambda}_{k A_j}} \xrightarrow{p} \chi^2(1), \quad \frac{\left( \hat{\lambda}_{k A_j} - \hat{\lambda}_{A_k} \right)}{1 - \hat{\lambda}_{k A_j}} \xrightarrow{p} 0$$

45
LR Test vs. COP

Given current $A$, the likelihood ratio (LR) test statistic of $H_0: X_j$ is irrelevant, vs. $H_1: X_j$ is relevant

$$2LR_j = -n \left( \sum_{k=1}^{K} \log (1 - \hat{\lambda}_{k}^{A+ j}) - \sum_{k=1}^{K} \log (1 - \hat{\lambda}_{k}^{A}) \right)$$

$$= n \sum_{k=1}^{K} \log \left( 1 + \frac{\hat{\lambda}_{k}^{A+ j} - \hat{\lambda}_{k}^{A}}{1 - \hat{\lambda}_{k}^{A+ j}} \right)$$

Under $H_0: X_j$ is irrelevant

$$COP_k^{A+ j} = n \frac{\lambda_{k}^{A+ j} - \hat{\lambda}_{k}^{A}}{1 - \hat{\lambda}_{k}^{A+ j}} \rightarrow_p \chi^2(1), \quad \frac{\hat{\lambda}_{k}^{A+ j} - \hat{\lambda}_{k}^{A}}{1 - \hat{\lambda}_{k}^{A+ j}} \rightarrow_p 0$$

$$2LR_j \rightarrow_p COP_{1:K}^{A+ j} = \sum_{k=1}^{K} COP_k^{A+ j} \rightarrow_p \chi^2(K)$$
Beyond the First-order

- $E(X_1|Y) = 0$
For $h = 1, \ldots, H$,

\[
X_A | Y \in S_h \sim N(\mu_h, \Sigma_h),
\]

\[
X_{A^c} | X_A, Y \in S_h \sim N(\alpha + \beta' X_A, \Sigma_0).
\]

If $\text{Cov}(X)$ is not degenerate, the \textit{minimum} set $A$ satisfying the above model is unique.

Likelihood ratio test for $H_0 : A = \mathcal{C}$ v.s. $H^*_1 : A = \mathcal{C} \cup \{j\}$,

\[
L^*_j|c = \frac{P_{M^*_1}(x|y)}{P_{M_0}(x|y)} = \frac{P_{M^*_1}(x_j|x_C, y)}{P_{M_0}(x_j|x_C, y)}
\]
Likelihood Ratio Test

Assume $\left[\hat{\sigma}^{(h)}_j\right]^2$ is the estimated variance by regressing $X_j$ on $X_C$ in slice $S_h$ and $\hat{\sigma}^2_j$ is the overall estimated variance. Then,

$$\hat{D}_{j|C} = \frac{2}{n} \log \left( L_{j|C}^* \right) = \log \hat{\sigma}^2_j - \sum_{h=1}^{H} \frac{n_h}{n} \log \left[ \hat{\sigma}^{(h)}_j \right]^2$$

When $A \subset C$ and $|C| = d$, $n\hat{D}_{j|C} \xrightarrow{d} \chi^2_{(H-1)(d+2)}$

$$\hat{D}_{j|C} \sim \log \left( 1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h} \right) - \sum_{h=1}^{H} \frac{n_h}{n} \log \left( \frac{Q_h / n_h}{\sum_{h=1}^{H} Q_h / n} \right)$$

where $Q_0 \sim \chi^2_{(H-1)(d+1)}$ and $Q_h \sim \chi^2_{n_h-d-1}$ ($h = 1, \ldots, H$) are mutually independent according to Cochran’s theorem.
As $n \to \infty$, we have

$$
\hat{D}_{j|c}^* \overset{\text{a.s.}}{\to} D_{j|c}^*
$$

$$
= \log \left( 1 + \frac{\text{Var}(M_j) - \text{Cov}(M_j, X_c) [\text{Var}(X_c)]^{-1} \text{Cov}(M_j, X_c)'}{\mathbb{E}(V_j)} \right)
$$

$$
+ \log \mathbb{E}(V_j) - \mathbb{E} \log (V_j)
$$

where $M_j = \mathbb{E}(X_j | X_c, \mathcal{S}(Y))$ and $V_j = \text{Var}(X_j | X_c, \mathcal{S}(Y))$.

$\hat{D}_{j|c}^*$ a.s. $\to 0$ if and only if $\mathbb{E}(X_j | X_c, \mathcal{S}(Y)) = \mathbb{E}(X_j | X_c)$ and $\text{Var}(X_j | X_c, \mathcal{S}(Y)) = \text{Var}(X_j | X_c)$

A stepwise procedure based on $\hat{D}_{j|c}^*$ is consistent when $p = O(n^\gamma)$ with $\gamma < 1/2$. ★
Example revisit

\( \text{Var} (X_1 | Y) \) depends on \( Y \).
$\mathbb{E}(X_2|X_1, Y)$ depends on $Y$. 
Sure independence screening (SIS) when \( p \gg n \)

- Independence screening: first independently selects variables based on their marginal relationships with the response and then applies refined methods in the second step.
- Rank predictors according to \( \hat{D}_{j|C}^* \) with \( C = \emptyset \):

\[
\hat{D}_{j|C=\emptyset}^* \xrightarrow{\text{a.s.}} \log \left( 1 + \frac{\text{Var} \left( \mathbb{E}(X_j|\mathbb{S}(Y)) \right)}{\mathbb{E} \left( \text{Var}(X_j|\mathbb{S}(Y)) \right)} \right) \\
+ \log \mathbb{E} \left( \text{Var}(X_j|\mathbb{S}(Y)) \right) - \mathbb{E} \log \left( \text{Var}(X_j|\mathbb{S}(Y)) \right)
\]

- The first \( n - 1 \) variables have a high probability to include the true predictors \( \mathcal{A} \) (almost surely under moderate conditions) even when \( \log(p) = O(n^\gamma) \) with \( \gamma < 1 \).
Siri: An interweaving strategy
Theoretical Properties

- Under moderate conditions, a stepwise procedure with forward selection and backward elimination is consistent when $p = O(n^{\gamma})$ with $\gamma < 1/2$.
- By choosing the threshold appropriately, the addition step will not stop selecting variables until all the true predictors have been included.
- Once all the true predictors have been included, all the redundant variables will be removed from the selected variables.
Conditions for consistency

- For \( j \in \mathcal{A} \), we have

\[
X_j | X_{\mathcal{A} - \{j\}}, Y \in S_h \sim \mathcal{N} \left( \alpha_j^{(h)} + \beta' X_{\mathcal{A} - \{j\}}, \sigma_j^2 \right).
\]

- Condition 1 (detectability): Let \( \alpha_j(Y) = \sum_h^{H} \alpha_j^{(h)} \mathbb{I}(Y \in S_h) \). There exist \( \xi > 0 \) and \( \kappa > 0 \) such that

\[
\text{Var} (\alpha_j(Y)) \geq \xi n^{-\kappa} \text{ for } j \in \mathcal{A}.
\]

- Condition 2 (dependency): The eigenvalues of \( \text{Var}(X) \) and \( \text{Var}(X|Y \in S_h) \) \((h = 1, \ldots, H)\) have positive lower and upper bounds.

- Condition 3 (dimensionality): \( \lim_{n \to \infty} (p) = \infty \) and \( p = o(n^\rho) \) with \( \rho > 0 \) and \( 2\rho + 2\kappa < 1 \).
Consistency of Stepwise Procedure

- Under Condition 1-3, as \( n \to \infty \), there exist constants \( c > 0 \) and \( \kappa \geq 0 \) such that
  \[
  \Pr \left( \min_{c : C^c \cap A \neq \emptyset} \max_{j \in C^c} \hat{D}_{j | c} \geq cn^{-\kappa} \right) \to 1, \text{ and }
  \]
  \[
  \Pr \left( \max_{c : C^c \cap A = \emptyset} \max_{j \in C^c} \hat{D}_{j | c} < Cn^{-\kappa} \right) \to 1 \text{ for any } C > 0.
  \]

- If we choose \( t_a = cn^{-\kappa} \) and \( t_d = (c/2)n^{-\kappa} \), then the addition step will not stop selecting variables until all the true predictors have been included.
Implementation Issues

- Under Condition 1-3, as $n \to \infty$, there exist constants $c > 0$ and $\kappa \geq 0$ such that

$$\Pr \left( \min_{c:C^c \cap A \neq \emptyset} \max_{j \in C^c} \hat{D}_j | C \geq cn^{-\kappa} \right) \to 1,$$

and

$$\Pr \left( \max_{c:C^c \cap A = \emptyset} \max_{j \in C^c} \hat{D}_j | C < Cn^{-\kappa} \right) \to 1$$

for any $C > 0$.

- If we choose $t_a = cn^{-\kappa}$ and $t_d = (c/2)n^{-\kappa}$, then the addition step will not stop selecting variables until all the true predictors have been included.

- Once all the true predictors have been included, all the redundant variables will be removed from the selected variables.
### Simulation I (linear)

$$Y = X_p \beta + \epsilon, \quad \epsilon \sim N(0,1), \quad Cov(X_i, X_j) = 0.5^{|i-j|}$$

$$n = 200, \quad p = 1000, \quad \beta = (3,1.5,1,1,2,1,0.9,1,1,1,0,...,0)^T$$

<table>
<thead>
<tr>
<th>Method</th>
<th>FP(0, 990)</th>
<th>FN(0, 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SIRI-C</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[CV minimizing classification error]</td>
<td>1.86 (0.222)</td>
<td>1.66 (0.117)</td>
</tr>
<tr>
<td><strong>SIRI-M</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[CV minimizing mean square error]</td>
<td>0.76 (0.120)</td>
<td>1.75 (0.114)</td>
</tr>
<tr>
<td><strong>COP</strong></td>
<td>1.62 (0.165)</td>
<td>1.67 (0.118)</td>
</tr>
<tr>
<td><strong>SIS-SCAD</strong></td>
<td>0.10 (0.030)</td>
<td>0.64 (0.069)</td>
</tr>
<tr>
<td><strong>LASSO</strong></td>
<td>5.40 (0.188)</td>
<td>0.00 (0.000)</td>
</tr>
</tbody>
</table>
Simulation II: hierarchical interactions

\[ \mathbf{X} \sim \text{MVN}(\mathbf{0}, \mathbf{I}_p) \] with \( n = 200, \ p = 1000, \) and

\[ Y = X_1 + X_1X_2 + X_1X_3 + 0.2\epsilon \]
Simulation III: non-hierarchical

$\mathbf{X} \sim \text{MVN}(\mathbf{0}, \mathbf{I}_p)$ with $n = 200$, $p = 1000$, and

$$Y = X_1 X_2 + X_1 X_3 + 0.2 \epsilon$$
Simulation IV: Non-multiplicative

\[ \mathbf{X} \sim \text{MVN}(\mathbf{0}, \mathbb{I}_p) \text{ with } n = 200, \ p = 1000, \text{ and} \]

\[ Y = \frac{X_1}{X_2 + X_3} + 0.2\epsilon \]
Simulation V (heteroscedastic, single index)

\[ Y = \frac{0.2 \epsilon}{1.5 + \sum_{j=1}^{8} X_j}, \quad X_p \sim \text{independent normal} \]

\[ n = 1000, \quad p = 1000 \]

<table>
<thead>
<tr>
<th>Method</th>
<th>FP(0, 992)</th>
<th>FN(0, 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRI-C</td>
<td>2.00 (0.163)</td>
<td>0.42 (0.138)</td>
</tr>
<tr>
<td>SIRI-M</td>
<td>0.43 (0.079)</td>
<td></td>
</tr>
<tr>
<td>COP</td>
<td>1.26 (0.128)</td>
<td>3.32 (0.192)</td>
</tr>
<tr>
<td>SIS-SCAD</td>
<td>3.23 (0.356)</td>
<td>8.00 (0.000)</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.64 (0.255)</td>
<td>8.00 (0.000)</td>
</tr>
</tbody>
</table>
Simulation VI (hub with linear effect)

\[ Y = X_1 + X_1 \times (X_2 + X_3) + 0.2 \epsilon, \quad X_p \sim \text{independent normal} \]
\[ n = 200, \quad p = 1000 \]

<table>
<thead>
<tr>
<th>Method</th>
<th>FP(0, 997)</th>
<th>FN(0, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRI-C</td>
<td>0.39 (0.115)</td>
<td>0.12 (0.046)</td>
</tr>
<tr>
<td>SIRI-M</td>
<td>0.03 (0.017)</td>
<td>0.04 (0.020)</td>
</tr>
<tr>
<td>SIS-SCAD-2</td>
<td>0.00 (0.000)</td>
<td>0.45 (0.068)</td>
</tr>
</tbody>
</table>
Simulation VII (three-way interaction)

\[ Y = X_1 \times X_2 \times X_3 + 0.2\epsilon, \quad X_p \sim \text{independent normal} \]

\[ n = 500, \quad p = 1000 \]
Bayesian Networks

Example: Printer Troubleshooting

26 variables
Instead of $2^{26}$ parameters we get
$99 = 17x1 + 1x2^1 + 2x2^2 + 3x2^3 + 3x2^4$

[Heckerman, 95]
Bayesian Network: \( \text{BN} = (\text{G, } \Theta) \)

- \( \text{G} \) - directed acyclic graph (DAG)
- \( \Theta \) - set of parameters in all conditional probability distributions (CPDs)

Compact representation of joint distribution in a product form (chain rule):
\[
P(S, C, B, X, D) = P(S) P(C|S) P(B|S) P(X|C,S) P(D|C,B)
\]

\[1 + 2 + 2 + 4 + 4 = 13 \text{ parameters instead of } 2^5 = 32\]
Learning BN structures

• Global approach (Score/likelihood based):
  • Posterior inference:
    \[ P(G \mid Data) \propto \int P(Data \mid \theta_G, G)p(\theta_G \mid G)p(G)d\theta_G \]
  • Or score-based criterion
    • AIC = \(-2 \log P(Data \mid \hat{\theta}_G, G) + 2p_G\)
    • BIC = \(-2 \log P(Data \mid \hat{\theta}_G, G) + \log(n)p_G\)
Learning structures

- Local approaches: using conditional independence statements as constraints.
  - Represented by “Inductive causality” (IC) algorithm due to Pearl (2000).

1. First, the skeleton of the network (undirected graph underlying the network structure) is learned by recursively testing the conditional independence between nodes.

2. Set all direction of the arcs that are part of a v-structure, which is a triplet of nodes incident on a converging connection $X_j \rightarrow X_i \leftarrow X_k$

3. Set the directions of the order arcs as needed to satisfy the acyclicity constraint.
Finding Markov blanket for each note using Growth-Shrink (GS) algorithm

- It is like a stepwise regression. For each note $X_i$, we treat it as the response variable and
  (a) gradually add variables that are predictive of $X_i$;
  (b) Backward removing those “redundant” $X_j$’s obtained from the growth phase.
Discussion

- Cross-validation to select the dimension and thresholds
- Back to full Bayesian model with dynamic slicing
  - We want to have flexibility in choosing slicing boundaries
  - Connection with Mutual-Information Criterion (MIC)
  - Many interesting possibilities
- Robustness to the distribution of predictors
Acknowledgment

Bo Jiang – who did all the work
Dr. Tingting Zhang, Dr. Wenxuan Zhong
Joseph K. Blitzstein