Structure Learning for Bayesian Networks over Labeled DAGs

Antti Hyttinen, Johan Pensar, Juha Kontinen, Jukka Corander

University of Helsinki
Helsinki Institute for Information Technology
Department of Computer Science
Department of Mathematics and Statistics

PGM2018, Prague, Czech Republic
\( X \perp\!\!\!\!\!\!\perp Y \mid C, Z = 0 \)

i.e. \( P(X\mid Y, C, Z = 0) = P(X\mid C, Z = 0) \)

but \( P(X\mid Y, C, Z = 1) \neq P(X\mid C, Z = 1) \) (possibly)
Context-specific Independence [Boutilier et al. ’96]

\[ X \perp \!\!\!\!\!\!\!\!\!\!\perp Y \mid C, Z = 0 \]

i.e. \( P(X|Y, C, Z = 0) = P(X|C, Z = 0) \)

but \( P(X|Y, C, Z = 1) \neq P(X|C, Z = 1) \) (possibly)

- A very natural independence restriction for any modelling task.
- For example:

\[
\begin{align*}
\text{INCOME} & \perp \!\!\!\!\!\!\!\!\!\!\perp \text{WEATHER} \mid \text{JOB} = \text{clerk} \\
\text{INCOME} & \not\perp \!\!\!\!\!\!\!\!\!\!\not\perp \text{WEATHER} \mid \text{JOB} = \text{farmer}
\end{align*}
\]
Context-specific Independence [Boutilier et al. ’96]

\[ X \perp \!\!\!\perp Y|C, Z = 0 \]

i.e. \( P(X|Y, C, Z = 0) = P(X|C, Z = 0) \)

but \( P(X|Y, C, Z = 1) \neq P(X|C, Z = 1) \) (possibly)

- A very natural independence restriction for any modelling task.
- For example:
  \[ \text{INCOME} \perp \!\!\!\perp \text{WEATHER}| \text{JOB} = \text{clerk} \]
  \[ \text{INCOME} \not\perp \!\!\!\not\perp \text{WEATHER}| \text{JOB} = \text{farmer} \]

- Alarm has several of these:
  \[ \text{HREKG} \perp \!\!\!\perp \text{CRRCAUTER}| \text{HR} = \text{LOW} \]
Motivation

- Can we orient causal edges based on CSIs in a principled way?

\[ Q \perp \perp X \]

\[ Y \]

AND \[ X \perp Y \mid Q = 0 \implies ? \]
Can we orient causal edges based on CSIs in a principled way?
What are good graphical models for understanding CSIs?
Motivation

Can we orient causal edges based on CSIs in a principled way?
What are good graphical models for understanding CSIs?
Can we get better causal or probabilistic models by using CSIs?
Contents

1 BNs over LDAGs

2 Separation Criteria

3 Constraint-based learning

4 Score-based learning

5 Conclusion
BNs over LDAGs
Bayesian Networks over Labeled DAGs [Pensar et al. 15]

\[
\begin{array}{c|cc}
P(X) & X = 0 & X = 1 \\
\hline
0.5 & 0.5 \\
\end{array}
\]

\[
\begin{array}{c|cc}
P(A) & A = 0 & A = 1 \\
\hline
0.5 & 0.5 \\
\end{array}
\]

\[
\begin{array}{c|cc}
P(Z|A, X) & Z = 0 & Z = 1 \\
\hline
AX = 00 & 0.1 & 0.9 \\
AX = 01 & 0.1 & 0.9 \\
AX = 10 & 0.5 & 0.5 \\
AX = 11 & 0.6 & 0.4 \\
\end{array}
\]

\[
\begin{array}{c|cc}
P(Y|A, Z) & Y = 0 & Y = 1 \\
\hline
AZ = 00 & 0.1 & 0.9 \\
AZ = 01 & 0.2 & 0.8 \\
AZ = 10 & 0.6 & 0.4 \\
AZ = 11 & 0.6 & 0.4 \\
\end{array}
\]
A label on an edge encodes contexts where the edge is absent. More formally:
A label on an edge encodes contexts where the edge is absent. More formally:

- Label on $X \rightarrow Z$ is a set of assignments to the other parents of $Z$: e.g. $A = 0$ on $X \rightarrow Z$. 
Bayesian Networks over Labeled DAGs [Pensar et al. 15]

<table>
<thead>
<tr>
<th>$P(X)$</th>
<th>$X = 0$</th>
<th>$X = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(A)$</th>
<th>$A = 0$</th>
<th>$A = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

| $P(Z|A,X)$ | $Z = 0$ | $Z = 1$ |
|----------|---------|---------|
| $AX = 00$ | 0.1     | 0.9     |
| $AX = 01$ | 0.1     | 0.9     |
| $AX = 10$ | 0.5     | 0.5     |
| $AX = 11$ | 0.6     | 0.4     |

- A label on an edge encodes contexts where the edge is absent. More formally:
- Label on $X \rightarrow Z$ is a set of assignments to the other parents of $Z$: e.g. $A = 0$ on $X \rightarrow Z$.
- Any assignment in a label denotes a local CSI: e.g. $X \perp Z| A = 0$. 
A label on an edge encodes contexts where the edge is absent. More formally:

- Label on $X \rightarrow Z$ is a set of assignments to the other parents of $Z$: e.g. $A = 0$ on $X \rightarrow Z$.
- Any assignment in a label denotes a local CSI: e.g. $X \perp Z | A = 0$.
- CPT has rows consistent with the assignment equal.
### Another Bayesian Network over a Labeled DAG

<table>
<thead>
<tr>
<th>$P(X)$</th>
<th>$X = 0$</th>
<th>$X = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(A)$</th>
<th>$A = 0$</th>
<th>$A = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

| $P(Z|A, X, Y)$ | $Z = 0$ | $Z = 1$ |
|---------------|---------|---------|
| $AXY = 000$   | 0.5     | 0.5     |
| $AXY = 001$   | 0.9     | 0.1     |
| $AXY = 010$   | 0.5     | 0.5     |
| $AXY = 011$   | 0.9     | 0.1     |
| $AXY = 100$   | 0.1     | 0.9     |
| $AXY = 101$   | 0.1     | 0.9     |
| $AXY = 110$   | 0.6     | 0.4     |
| $AXY = 111$   | 0.6     | 0.4     |

| $P(Y|A)$ | $Y = 0$ | $Y = 1$ |
|----------|---------|---------|
| $A = 0$  | 0.1     | 0.9     |
| $A = 1$  | 0.6     | 0.4     |

- **Local CSIs:** $X \perp Z| AY = 00$, $X \perp Z| AY = 01$, $Y \perp Z| AX = 10$, $Y \perp Z| AX = 11$
Alternative modelling strategies [Koller & Friedman, ch. 5]:

- Decision tree-based CPTs (subsumed in the binary case)
- Rule-CPTs
- Noisy-ORs, logistic models, etc.
Modelling local structure in BN CPTs

Alternative modelling strategies [Koller & Friedman, ch. 5]:

- Decision tree-based CPTs (subsumed in the binary case)
- Rule-CPTs
- Noisy-ORs, logistic models, etc.

LDAGs [Pensar et al. '15]:

- Allow for developing theory using the labels.
- Markov equivalence defined based on the labels.
- Visual representation of CSIs in a single structure.
Separation Criteria
CSI-separation of [Boutilier et al. 96] for LDAGs

- In a context $S = s$ specific DAG of an LDAG edges with labels consistent with $S = s$ are removed.
CSI-separation of [Boutilier et al. 96] for LDAGs

Original LDAG

\[
\begin{align*}
\text{Original LDAG} & \Rightarrow \\
A & \downarrow \\
X & \quad Y \\
A = 0 & \quad A = 1 \\
Z & \\
\end{align*}
\]

context \( A = 1 \) specific DAG

\[
\begin{align*}
\text{context } A = 1 \text{ specific DAG} & \Rightarrow \\
X & \quad Y \\
\Rightarrow & \\
X \perp_{CSI} Y \mid A = 1
\end{align*}
\]

- In a context \( S = s \) specific DAG of an LDAG edges with labels consistent with \( S = s \) are removed.

- \( X \) and \( Y \) are CSI-separated given \( C, S = s \), iff \( X \) and \( Y \) are d-separated given \( C, S \) in the context \( S = s \) specific DAG.
CSI-separation of [Boutilier et al. 96] for LDAGs

In a context $S = s$ specific DAG of an LDAG, edges with labels consistent with $S = s$ are removed.

- $X$ and $Y$ are CSI-separated given $C, S = s$, iff $X$ and $Y$ are d-separated given $C, S$ in the context $S = s$ specific DAG.

- CSI-separation is sound and it subsumes d-separation.
CSI-separation of [Boutilier et al. 96] for LDAGs

- In a context $S = s$ specific DAG of an LDAG edges with labels consistent with $S = s$ are removed.

- $X$ and $Y$ are CSI-separated given $C, S = s$, iff $X$ and $Y$ are d-separated given $C, S$ in the context $S = s$ specific DAG.

- CSI-separation is sound and it subsumes d-separation.

- But CSI-sep. is incomplete: $X \perp \perp Y$! NP-hard!
New Necessary Separation Criterion for LDAGs

<table>
<thead>
<tr>
<th>Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>For $X \perp \perp Y \mid C$, $S = v[S]$ to be implied by an LDAG over $V$, $X, Y$ have to be a $d$-separated given $C, S$ in all context $V = v$ specific DAGs.</td>
</tr>
</tbody>
</table>
New Necessary Separation Criterion for LDAGs

Theorem

For $X \perp\!\!\!\!\!\!\perp Y \mid C, S = v[S]$ to be implied by an LDAG over $V$
$X, Y$ have to be d-separated given $C, S$
in all context $V = v$ specific DAGs.

- E.g. on right $X, Y$ are d-connected
given $Z$ when $Q = 0, R = 0$,
thus there are parameters such that
$X \not\!\!\!\!\!\!\perp\!\!\!\!\!\!\perp Y \mid Z$. 

\[ \begin{array}{c}
Q \to R \\
X \to Z \\
Z \to Y
\end{array} \]

\[ \begin{array}{c}
R = 1 \\
Q = 1
\end{array} \]
New Necessary Separation Criterion for LDAGs

Theorem

For $X \perp \perp Y \mid C$, $S = v[S]$ to be implied by an LDAG over $V$

$X$, $Y$ have to be $d$-separated given $C, S$

in all context $V = v$ specific DAGs.

- E.g. on right $X$, $Y$ are $d$-connected
  given $Z$ when $Q = 0, R = 0$,
  thus there are parameters such that $X \not\perp \perp Y \mid Z$.

- If nodes are $d$-separated in all context $V = v$ specific DAGs,
  but not CSI-separated, they may be independent or dependent.
New Necessary Separation Criterion for LDAGs

Theorem

For $X \perp \perp Y|C$, $S = v[S]$ to be implied by an LDAG over $V$
$X$, $Y$ have to be a $d$-separated given $C$, $S$
in all context $V = v$ specific DAGs.

- E.g. on right $X$, $Y$ are $d$-connected
given $Z$ when $Q = 0$, $R = 0$,
thus there are parameters such that
$X \not\perp \not\perp Y|Z$.

- If nodes are $d$-separated in all context $V = v$ specific DAGs,
but not CSI-separated, they may be independent or dependent.

- In the following we assume faithfulness w.r.t. to the theorem.
LDAGs are Markov equivalent iff all their context $V = v$ specific DAGs are Markov equivalent [Pensar et al. 15].
Markov Equivalence for LDAGs

- LDAGs are Markov equivalent iff all their context $V = v$ specific DAGs are Markov equivalent [Pensar et al. 15].

- **LDAG-colliders**: $X \rightarrow Z \leftarrow A$ without $X \rightarrow A$ in some context $V = v$ specific DAG

- **LDAG-non-colliders**: $Z \leftarrow A \rightarrow Y$ without $Z \leftarrow Y$ in some context $V = v$ specific DAG
Markov Equivalence for LDAGs

- LDAGs are Markov equivalent iff all their context $V = v$ specific DAGs are Markov equivalent [Pensar et al. 15].
- **LDAG-colliders**: $X \rightarrow Z \leftarrow A$ without $X \rightarrow A$ in some context $V = v$ specific DAG
- **LDAG-non-colliders**: $Z \rightarrow A \rightarrow Y$ without $Z \rightarrow Y$ in some context $V = v$ specific DAG
- Markov equivalent LDAGs share them: $X \rightarrow Z \rightarrow Y$ is neither.
Constraint-based learning
PC of Spirtes et al.

1. Skeleton search: Try to find a separating set $S$ such that $X \perp \perp Y \mid S$.
2. Orient colliders: $X \to Z \leftarrow Y$ if $Z \notin S$.
3. Run further orientation rules to make sure no cycles or new colliders are possible.

PC produces wrong orientation in the presence of CSIs!
PC of Spirtes et al.

1. Skeleton search: Try to find a separating set $S$ such that $X \perp \perp Y \mid S$.
2. Orient colliders: $X \rightarrow Z \leftarrow Y$ if $Z \notin S$.
3. Run further orientation rules to make sure no cycles or new colliders are possible.

![Graphs]

True graph

PC skeleton

PC result

PC produces wrong orientation in the presence of CSIs!
PC of Spirtes et al.

1. Skeleton search: Try to find a separating set $S$ such that $X \perp \perp Y | S$.
2. Orient colliders: $X \to Z \leftarrow Y$ if $Z \not\in S$.
3. Run further orientation rules to make sure no cycles or new colliders are possible.

PC produces wrong orientation in the presence of CSIs!
LPC Skeleton Search

• Instead, we search for **separating contexts** $S = s$, s.t. $X \perp \perp Y \mid S = s$.
• Delete edges if $X \perp \perp Y \mid S = s$ for all $s$.
• Otherwise record the separating contexts on the edge.
• Instead, we search for separating contexts \( S = s \), s.t. \( X \perp \perp Y \mid S = s \).

• Delete edges if \( X \perp \perp Y \mid S = s \) for all \( s \).

• Otherwise record the separating contexts on the edge.

\[
\begin{align*}
&X \perp A \\
&X \perp Y \\
&X \perp Z \mid A = 0 \\
&Z \perp Y \mid A = 1 \\
&X \perp Z \mid AY = 00 \\
&Z \perp Y \mid AY = 10 \\
&X \perp Z \mid AY = 01 \\
&Z \perp Y \mid AY = 11 \\
\end{align*}
\]

True graph

CSIs

LPC skeleton
In the paper we give technical conditions for detecting LDAG-(non-)colliders from the LPC skeleton result.
In the paper we give technical conditions for detecting LDAG-(non-)colliders from the LPC skeleton result.

1. LDAG-colliders can be oriented: e.g. $X \rightarrow Z \leftarrow A$.
2. LDAG-non-colliders are used in further orientation with modified PC rules [Meek '95].
In the paper we give technical conditions for detecting LDAG-(non-)colliders from the LPC skeleton result.

LDAG-colliders can be oriented: e.g. $X \rightarrow Z \leftarrow A$.

LDAG-non-colliders are used in further orientation with modified PC rules [Meek '95].
In the paper we give technical conditions for detecting LDAG-(non-)colliders from the LPC skeleton result.

LDAG-colliders can be oriented: e.g. $X \rightarrow Z \leftarrow A$.

LDAG-non-colliders are used in further orientation with modified PC rules [Meek '95].

LPC is conjectured to be orientation complete.
Simulations: Orientation Accuracy

10-node binary LDAGs, 300 models, over the true distribution.

<table>
<thead>
<tr>
<th>algo</th>
<th>av. degree</th>
<th>label prob.</th>
<th>edges found</th>
<th>corr. oriented</th>
<th>reversed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>2.99</td>
<td>0%</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
<tr>
<td>cPC</td>
<td>2.99</td>
<td>0%</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
<tr>
<td>LPC</td>
<td>2.99</td>
<td>0%</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
</tbody>
</table>

Without CSIs due to labels, algorithms work similarly.

With CSIs due to labels, PC makes orientation errors.

• cPC does not but orients less.
• LPC orients more and all orientations are correct.
Simulations: Orientation Accuracy

10-node binary LDAGs, 300 models, over the true distribution.

<table>
<thead>
<tr>
<th>algo</th>
<th>av. degree</th>
<th>label prob.</th>
<th>edges found</th>
<th>corr. oriented</th>
<th>reversed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>2.99</td>
<td>0 %</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
<tr>
<td>cPC</td>
<td>2.99</td>
<td>0 %</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
<tr>
<td>LPC</td>
<td>2.99</td>
<td>0 %</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
</tbody>
</table>

- Without CSIs due to labels, algorithms work similarly.
10-node binary LDAGs, 300 models, over the true distribution.

<table>
<thead>
<tr>
<th>algo</th>
<th>av. degree</th>
<th>label prob.</th>
<th>edges found</th>
<th>corr. oriented</th>
<th>reversed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>2.99</td>
<td>0 %</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
<tr>
<td>cPC</td>
<td>2.99</td>
<td>0 %</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
<tr>
<td>LPC</td>
<td>2.99</td>
<td>0 %</td>
<td>4481</td>
<td>3498</td>
<td>0</td>
</tr>
</tbody>
</table>

- Without CSIs due to labels, algorithms work similarly.

<table>
<thead>
<tr>
<th>algo</th>
<th>av. degree</th>
<th>label prob.</th>
<th>edges found</th>
<th>corr. oriented</th>
<th>reversed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>2.18</td>
<td>50 %</td>
<td>3276</td>
<td><strong>2243</strong></td>
<td><strong>103</strong></td>
</tr>
<tr>
<td>cPC</td>
<td>2.18</td>
<td>50 %</td>
<td>3276</td>
<td><strong>2285</strong></td>
<td>0</td>
</tr>
<tr>
<td>LPC</td>
<td>2.18</td>
<td>50 %</td>
<td>3276</td>
<td><strong>2319</strong></td>
<td>0</td>
</tr>
</tbody>
</table>

- With CSIs due to labels, PC makes orientation errors.
- cPC does not but orients less.
- LPC orients more and all orientations are correct.
Score-based learning
Maximizing BIC [Chickering ’97]:

$$\max_G \sum_{X \in V} s(X, \text{pa}_G(X)),$$

$$s(X, \text{pa}_G(X)) = \max_{\text{LABELS}} s(X, \text{pa}_G(X), \text{LABELS})$$
BIC for LDAGs

• Maximizing BIC [Chickering '97]:

$$\max_G \sum_{X \in V} s(X, \text{pa}_G(X)),$$

$$s(X, \text{pa}_G(X)) = \max_{\text{LABELS}} s(X, \text{pa}_G(X), \text{LABELS})$$

• LABELS imply a partition of rows:

\[
\begin{array}{c|cc}
Z = 0 & Z = 1 \\
\hline
AX = 00 & \theta_1 & 1 - \theta_1 \\
AX = 01 & \theta_1 & 1 - \theta_1 \\
AX = 10 & \theta_2 & 1 - \theta_2 \\
AX = 11 & \theta_3 & 1 - \theta_3 \\
\end{array}
\]

$$s(X, \text{pa}_G(X), \text{LABELS}) = L - R \cdot \log N/2$$

$L$ is max. likelihood, $R$ number of parts, both w.r.t. LABELS.
BIC for LDAGs

- Maximizing BIC [Chickering '97]:
  \[
  \max_G \sum_{X \in V} s(X, \text{pa}_G(X)), \\
  s(X, \text{pa}_G(X)) = \max_{\text{LABELS}} s(X, \text{pa}_G(X), \text{LABELS})
  \]

- LABELS imply a partition of rows:
  \[
  \begin{align*}
  P(Z|A, X) &| Z = 0 & Z = 1 \\
  AX = 00 & \theta_1 & 1 - \theta_1 \\
  AX = 01 & \theta_1 & 1 - \theta_1 \\
  AX = 10 & \theta_2 & 1 - \theta_2 \\
  AX = 11 & \theta_3 & 1 - \theta_3
  \end{align*}
  \]
  \[
  \Larr \{ \{1, 2\}, \{3\}, \{4\} \}
  \]
  \[
  s(X, \text{pa}_G(X), \text{LABELS}) = L - R \cdot \log N/2
  \]
  \[L\text{ is max. likelihood, } R\text{ number of parts, both w.r.t. LABELS.}\]
- For 4 binary parents, 27 million different label structures.
Branching ...

\[
\begin{align*}
\{ \{1\}, \{2\}, \{3\}, \{4\} \} & \rightarrow \{ \{1,2\}, \{3\}, \{4\} \} & \rightarrow \{ \{1,2,3\}, \{4\} \} & \rightarrow \{ \{1,2,4\}, \{3\} \} \\
\{ \{1\}, \{2\}, \{3\}, \{4\} \} & \rightarrow \{ \{1,2\}, \{3\}, \{4\} \} & \rightarrow \{ \{1,2,3\}, \{4\} \} & \rightarrow \{ \{1,2\}, \{3,4\} \}
\end{align*}
\]

- Search over partitions of rows from complex towards simpler.
Search over partitions of rows from complex towards simpler.

Keep a set of parts fixed (in red).

Combine the first unfixed part to the fixed parts to avoid visiting the same partitions more than once (symmetry breaking).
...and Bounding

\[
\{ \{1\}, \{2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2,3\}, \{4\} \} \rightarrow \{ \{1,2,4\}, \{3\} \}
\]

- **Upper bound** for partitions further in the branch:

\[
L - f \cdot \log N/2
\]

Here \(L\) is the current likelihood, \(f\) is the number of fixed parts.
...and Bounding

\[
\{ \{1\}, \{2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2,3\}, \{4\} \} \rightarrow \{ \{1,2,4\}, \{3\} \} \rightarrow \{ \{1,2\}, \{3,4\} \}
\]

- **Upper bound** for partitions further in the branch:

  \[
  L - f \cdot \log N/2
  \]

  Here \(L\) is the current likelihood, \(f\) is the number of fixed parts.

- **Initial best:** best solution to the subsets of parents.
... and Bounding

\[
\{\{1\}, \{2\}, \{3\}, \{4\}\} \rightarrow \{\{1,2\}, \{3\}, \{4\}\} \rightarrow \{\{1,2,3\}, \{4\}\}
\]

- **Upper bound** for partitions further in the branch:

\[L - f \cdot \log N/2\]

Here \(L\) is the current likelihood, \(f\) is the number of fixed parts.

- **Initial best**: best solution to the subsets of parents.

- **LDAG consistency check** whenever a new best found.
...and Bounding

\[
\{ \{1\}, \{2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2,3\}, \{4\} \} \rightarrow \{ \{1,2,4\}, \{3\} \}
\]

- **Upper bound** for partitions further in the branch:

\[
L - f \cdot \log \frac{N}{2}
\]

Here \(L\) is the current likelihood, \(f\) is the number of fixed parts.

- **Initial best**: best solution to the subsets of parents.
- **LDAG consistency check** whenever a new best found.
- Scales up to 4 parents.
...and Bounding

\[
\{ \{1\}, \{2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2\}, \{3\}, \{4\} \} \rightarrow \{ \{1,2,3\}, \{4\} \} \rightarrow \{ \{1,2,4\}, \{3\} \}
\]

\[
\rightarrow \ldots \rightarrow \{ \{1,2\}, \{3,4\} \}
\]

- **Upper bound** for partitions further in the branch:

\[
L - f \cdot \log N/2
\]

Here \( L \) is the current likelihood, \( f \) is the number of fixed parts.

- **Initial best**: best solution to the subsets of parents.
- **LDAG consistency check** whenever a new best found.
- Scales up to 4 parents.
- Finally: maximization over the local scores by **Gobnilp**.
• An extra edge does not always increase the BIC penalty:

\[ P(Z|X) \]

<table>
<thead>
<tr>
<th>( X = 0 )</th>
<th>( Z = 0 )</th>
<th>( Z = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

\[ P(Z|X, Y) \]

<table>
<thead>
<tr>
<th>( XY = 00 )</th>
<th>( Z = 0 )</th>
<th>( Z = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

\[ XY = 01 \]  
\[ XY = 10 \]  
\[ XY = 11 \]  

\[ \Rightarrow \]

\[ P(Z|X, Y) \]

<table>
<thead>
<tr>
<th>( XY = 00 )</th>
<th>( Z = 0 )</th>
<th>( Z = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.6</td>
<td></td>
</tr>
</tbody>
</table>

\[ XY = 01 \]  
\[ XY = 10 \]  
\[ XY = 11 \]  

\[ 0.7 \]  
\[ 0.3 \]
• An extra edge does not always increase the BIC penalty:

\[
P(Z|X) \begin{array}{c|cc}
X = 0 & Z = 0 & Z = 1 \\
0.4 & 0.6 & \\
X = 1 & 0.6 & 0.4
\end{array}
\Rightarrow
\[
P(Z|X, Y) \begin{array}{c|cc}
X Y & Z = 0 & Z = 1 \\
XY = 00 & 0.4 & 0.6 \\
XY = 01 & 0.4 & 0.6 \\
XY = 10 & 0.4 & 0.6 \\
XY = 11 & 0.7 & 0.3
\end{array}
\]

• **Strong Score Pruning** Delete a local score if it is not better than for a subset by a margin controlled by \( t \).

• **Mixed BIC Penalty** Penalize by

\[
a \cdot \text{LDAG-based BIC} + b \cdot \text{DAG-based BIC}.
\]

• **LDAG over Optimal DAG Skeleton** Only orient with the LDAG-based BIC score.
10-node binary LDAGs, 0.5 label probability. At most 3 parents.

- LDAG-based BIC overfits considerably (red).

With strong score pruning LDAG is better than a DAG (yellow vs. purple).
With more samples DAGs catch up but still keep CSIs hidden.
10-node binary LDAGs, 0.5 label probability. At most 3 parents.

- LDAG-based BIC overfits considerably (red).
- With strong score pruning LDAG is better than a DAG (yellow vs. purple).
10-node binary LDAGs, 0.5 label probability. At most 3 parents.

- LDAG-based BIC overfits considerably (red).
- With strong score pruning LDAG is better than a DAG (yellow vs. purple).
- With more samples DAGs catch up but still keep CSIs hidden.
• Structure learning for labeled DAGs.
Conclusion

- Structure learning for labeled DAGs.
- Principled orientation of causal edges using CSIs with LPC:
  - Based on separation criteria and Markov equivalence of LDAGs.
  - More orientations more correctly than PC when CSIs present.

![Diagram](image-url)
Structure learning for labeled DAGs.

Principled orientation of causal edges using CSIs with LPC:
- Based on separation criteria and Markov equivalence of LDAGs.
- More orientations more correctly than PC when CSIs present.

Better probabilistic models with score-based discovery:
- Using the LDAG-based BIC score.
- A Branch and Bound for local score calculation.
- Strong score pruning to avoid overfitting.
• Structure learning for labeled DAGs.
• Principled orientation of causal edges using CSIs with LPC:
  • Based on separation criteria and Markov equivalence of LDAGs.
  • More orientations more correctly than PC when CSIs present.
• Better probabilistic models with score-based discovery:
  • Using the LDAG-based BIC score.
  • A Branch and Bound for local score calculation.
  • Strong score pruning to avoid overfitting.
• CSIs are common and powerful but discovering them in sample data can be quite challenging!