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

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)

Context-specific Independence [Boutilier et al. '96]

$$X \perp\!\!\!\perp Y | C, Z = 0$$

$$\text{i.e. } P(X|Y, C, Z = 0) = P(X|C, Z = 0)$$

$$\text{but } P(X|Y, C, Z = 1) \neq P(X|C, Z = 1) \text{ (possibly)}$$

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

$$\text{INCOME} \perp\!\!\!\perp \text{WEATHER} |\ \text{JOB} = \text{clerk}$$

$$\text{INCOME} \not\perp\!\!\!\perp \text{WEATHER} |\ \text{JOB} = \text{farmer}$$

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:

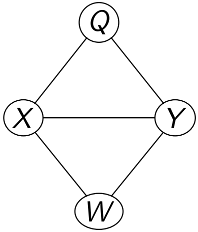
$$\text{INCOME} \perp\!\!\!\perp \text{WEATHER}|\ \text{JOB} = \text{clerk}$$

$$\text{INCOME} \not\perp\!\!\!\perp \text{WEATHER}|\ \text{JOB} = \text{farmer}$$

- Alarm has several of these:

$$\text{HREKG} \perp\!\!\!\perp \text{CRRCAUTER}|\ \text{HR} = \text{LOW}$$

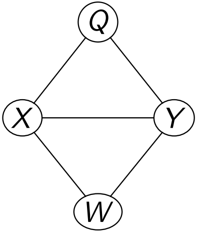
Motivation



AND $X \perp\!\!\!\perp Y | \; Q = 0 \quad \Rightarrow \quad ?$

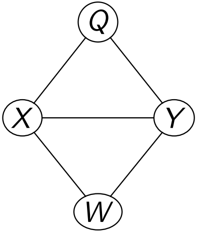
- Can we orient causal edges based on CSIs in a principled way?

Motivation



AND $X \perp\!\!\!\perp Y \mid Q = 0 \quad \Rightarrow \quad ?$

- Can we orient causal edges based on CSIs in a principled way?
- What are good graphical models for understanding CSIs?



AND $\quad X \perp\!\!\!\perp Y \mid Q = 0 \quad \Rightarrow \quad ?$

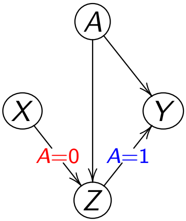
- Can we orient causal edges based on CSIs in a principled way?
- What are good graphical models for understanding CSIs?
- Can we get better causal or probabilistic models by using CSIs?

Contents

BNs over LDAGs

Bayesian Networks over Labeled DAGs [Pensar et al. 15]

$P(X)$	$X=0$	$X=1$
	0.5	0.5



$P(A)$	$A=0$	$A=1$
	0.5	0.5

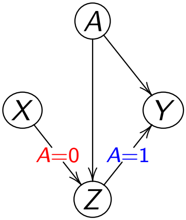
$P(Z\|A,X)$	$Z=0$	$Z=1$
$AX=00$	0.1	0.9
$AX=01$	0.1	0.9
$AX=10$	0.5	0.5
$AX=11$	0.6	0.4

$P(Y\|A,Z)$	$Y=0$	$Y=1$
$AZ=00$	0.1	0.9
$AZ=01$	0.2	0.8
$AZ=10$	0.6	0.4
$AZ=11$	0.6	0.4

Bayesian Networks over Labeled DAGs [Pensar et al. 15]

$P(X)$	$X=0$	$X=1$
	0.5	0.5

$P(A)$	$A=0$	$A=1$
	0.5	0.5



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

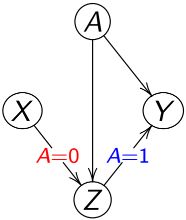
$P(Y\|A,Z)$	$Y=0$	$Y=1$
$AZ=00$	0.1	0.9
$AZ=01$	0.2	0.8
$AZ=10$	0.6	0.4
$AZ=11$	0.6	0.4

- **A label on an edge encodes contexts where the edge is absent. More formally:**

Bayesian Networks over Labeled DAGs [Pensar et al. 15]

$P(X)$	$X=0$	$X=1$
	0.5	0.5

$P(A)$	$A=0$	$A=1$
	0.5	0.5



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

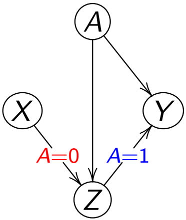
$P(Y\|A,Z)$	$Y=0$	$Y=1$
$AZ=00$	0.1	0.9
$AZ=01$	0.2	0.8
$AZ=10$	0.6	0.4
$AZ=11$	0.6	0.4

- **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 $X \rightarrow Z$.

Bayesian Networks over Labeled DAGs [Pensar et al. 15]

$P(X)$	$X=0$	$X=1$
	0.5	0.5

$P(A)$	$A=0$	$A=1$
	0.5	0.5



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

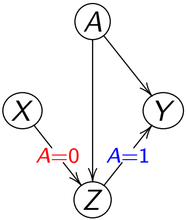
$P(Y\|A,Z)$	$Y=0$	$Y=1$
$AZ=00$	0.1	0.9
$AZ=01$	0.2	0.8
$AZ=10$	0.6	0.4
$AZ=11$	0.6	0.4

- **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. $X \perp\!\!\!\perp Z |\ A=0$.

Bayesian Networks over Labeled DAGs [Pensar et al. 15]

$P(X)$	$X=0$	$X=1$
	0.5	0.5

$P(A)$	$A=0$	$A=1$
	0.5	0.5



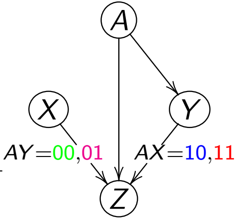
$P(Z\|A,X)$	$Z=0$	$Z=1$
$AX=00$	0.1	0.9
$AX=01$	0.1	0.9
$AX=10$	0.5	0.5
$AX=11$	0.6	0.4

$P(Y\|A,Z)$	$Y=0$	$Y=1$
$AZ=00$	0.1	0.9
$AZ=01$	0.2	0.8
$AZ=10$	0.6	0.4
$AZ=11$	0.6	0.4

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

Another Bayesian Network over a Labeled DAG

$P(X)$	$X=0$	$X=1$
	0.5	0.5



$P(A)$	$A=0$	$A=1$
	0.5	0.5

$P(Z\|A,X,Y)$	$Z=0$	$Z=1$
$AXY=000$	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: $X \perp\!\!\!\perp Z \mid AY=00$, $X \perp\!\!\!\perp Z \mid AY=01$, $Y \perp\!\!\!\perp Z \mid AX=10$, $Y \perp\!\!\!\perp Z \mid AX=11$

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.

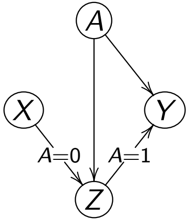
Modelling local structure in BN CPTs

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

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

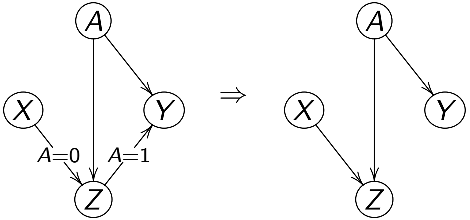
LDAGs [Pensar et al. '15]:

- 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

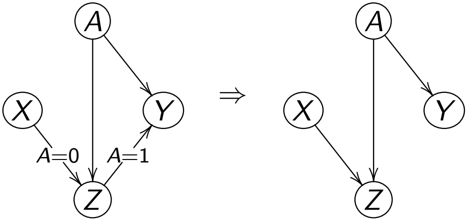


Original LDAG

context $A = 1$ specific DAG

- 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



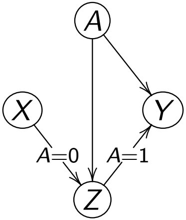
Original LDAG — context $A = 1$ specific DAG

$\Rightarrow \quad X \perp_{CSI} Y | \; A = 1$

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

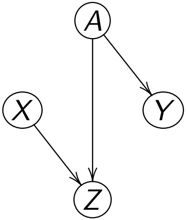
- X and Y are **CSI-separated** given $C, S = s$, iff X and Y are d-separated given C, S in the context $S = s$ specific DAG.

CSI-separation of [Boutilier et al. 96] for LDAGs



Original LDAG

$\Rightarrow$



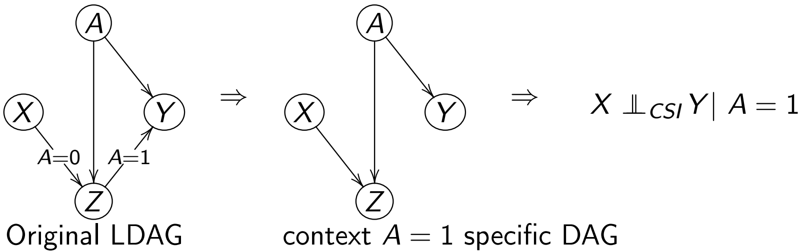
context $A = 1$ specific DAG

$\Rightarrow$

$$X \perp_{CSI} Y | \; A = 1$$

- 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

Theorem

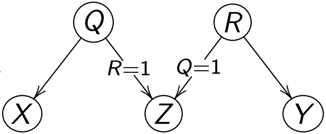
For $X \perp\!\!\!\perp Y | C, S = v[S]$ to be implied by an LDAG over V
X, Y have to be a d-separated given C, S
in all context $V = v$ specific DAGs.

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 $X \not\perp\!\!\!\perp Y|\, Z$.

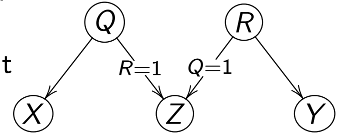


New Necessary Separation Criterion for LDAGs

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 $X \not\perp\!\!\!\perp Y | Z$.



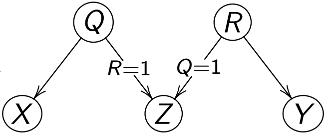
- If nodes are d-separated in all context $V = v$ specific DAGs, but not CSI-separated, they may be independent or dependent.

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.

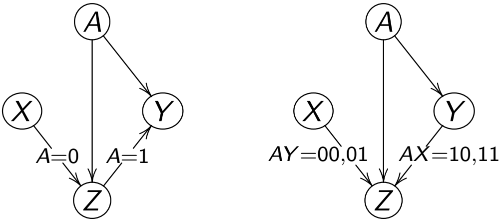
- E.g. on right X, Y are d-connected given Z when $Q = 0, R = 0$, thus there are parameters such that $X \not\!\perp\!\!\!\perp Y | Z$.



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

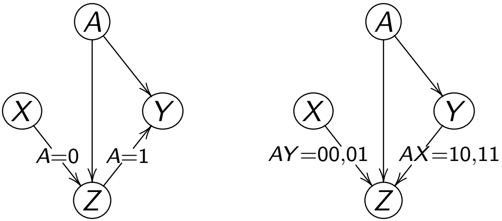
- In the following we assume faithfulness w.r.t. to the theorem.

Markov Equivalence for LDAGs



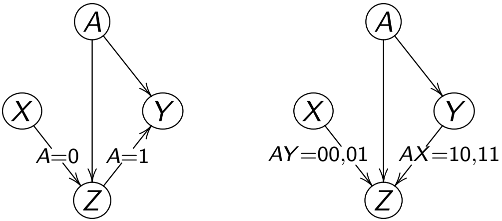
- LDAGs are Markov equivalent iff
  all their context $V = v$ specific DAGs are Markov equivalent
  [Pensar et al. 15].

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**: $X \to Z \leftarrow A$ without $X - A$ in some context $V = v$ specific DAG
- **LDAG-non-colliders**: $Z - A - Y$ without $Z - Y$ 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 $X - A$ in some context $V = v$ specific DAG
- **LDAG-non-colliders**: $Z - A - Y$ without $Z - Y$ in some context $V = v$ specific DAG
- Markov equivalent LDAGs share them: $X - Z - Y$ is neither.

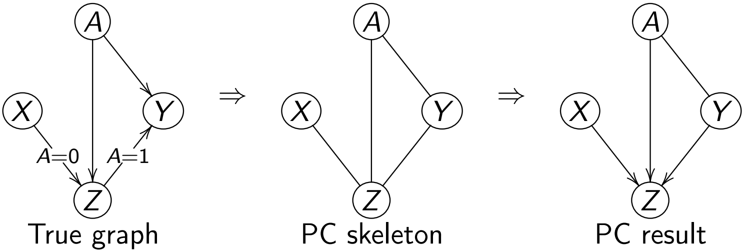
Constraint-based learning

PC of Spirtes et al.

1. Skeleton search: Try to find a separating set S such that $X \perp\!\!\!\perp Y | S$.
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.

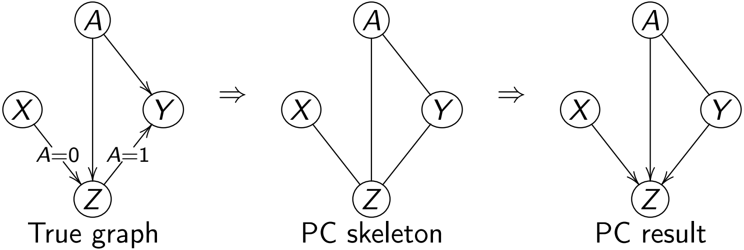
PC of Spirtes et al.

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



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: $X \rightarrow Z \leftarrow Y$ if $Z \notin S$.
3. Run further orientation rules to make sure no cycles or new colliders are possible.

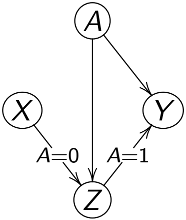


PC produces wrong orientation in the presence of CSIs!

LPC Skeleton Search

- Instead, we search for **separating contexts** $S = s$, s.t. $X \perp\!\!\!\perp Y|\, S = s$.
- Delete edges if $X \perp\!\!\!\perp Y|\, S = s$ for all s.
- Otherwise record the separating contexts on the edge.

- Instead, we search for **separating contexts** $S = s$, s.t. $X \perp\!\!\!\perp Y | \; S = s$.
- Delete edges if $X \perp\!\!\!\perp Y | \; S = s$ for all s.
- Otherwise record the separating contexts on the edge.



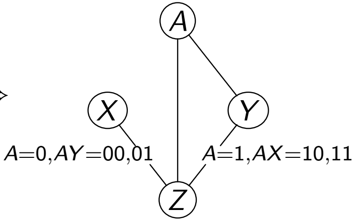
True graph

$\Rightarrow$

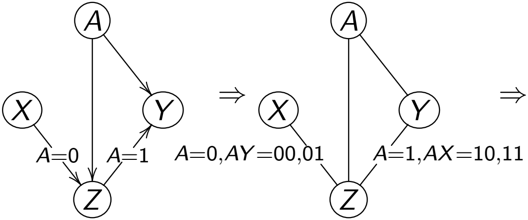
$$X \perp\!\!\!\perp A$$
$$X \perp\!\!\!\perp Y$$
$$X \perp\!\!\!\perp Z | \; A = 0$$
$$Z \perp\!\!\!\perp Y | \; A = 1$$
$$X \perp\!\!\!\perp Z | AY = 00$$
$$Z \perp\!\!\!\perp Y | AY = 10$$
$$X \perp\!\!\!\perp Z | AY = 01$$
$$Z \perp\!\!\!\perp Y | AY = 11$$

CSIs

$\Rightarrow$



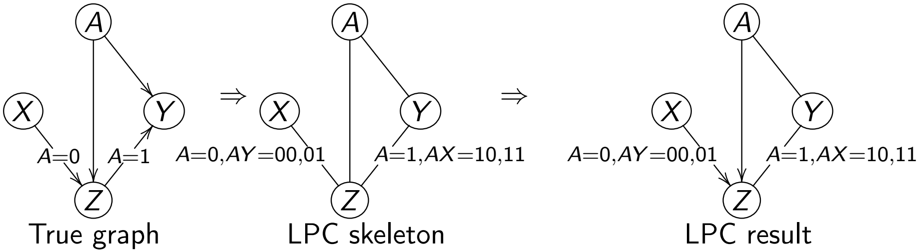
LPC skeleton



True graph LPC skeleton

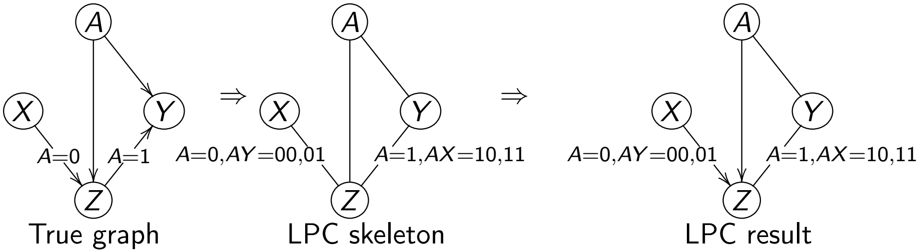
- In the paper we give technical conditions for detecting LDAG-(non-)colliders from the LPC skeleton result.

LPC Orientation



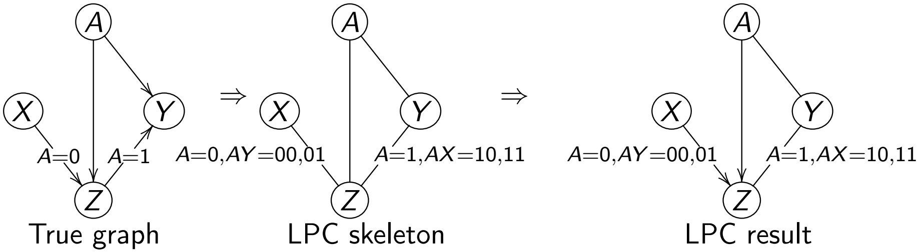
- In the paper we give technical conditions for detecting LDAG-(non-)colliders from the LPC skeleton result.
- LDAG-colliders can be oriented: e.g. $X \rightarrow Z \leftarrow A$.

LPC Orientation



- 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 Orientation



- 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.

Simulations: Orientation Accuracy

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

algo	av. degree	label prob.	edges found	corr. oriented	reversed
PC	2.99	0 %	4481	3498	0
cPC	2.99	0 %	4481	3498	0
LPC	2.99	0 %	4481	3498	0

- Without CSIs due to labels, algorithms work similarly.

Simulations: Orientation Accuracy

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

algo	av. degree	label prob.	edges found	corr. oriented	reversed
PC	2.99	0 %	4481	3498	0
cPC	2.99	0 %	4481	3498	0
LPC	2.99	0 %	4481	3498	0

- Without CSIs due to labels, algorithms work similarly.

algo	av. degree	label prob.	edges found	corr. oriented	reversed
PC	2.18	50 %	3276	**2243**	**103**
cPC	2.18	50 %	3276	**2285**	**0**
LPC	2.18	50 %	3276	**2319**	**0**

- 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

BIC for LDAGs

- Maximizing BIC [Chickering '97]:

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

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

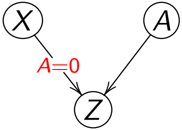
BIC for LDAGs

- Maximizing BIC [Chickering '97]:

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

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

- LABELS imply a partition of rows:



$\Leftrightarrow$

$P(Z \mid A, X)$	$Z = 0$	$Z = 1$
$AX = 00$	θ_1	$1 - \theta_1$
$AX = 01$	θ_1	$1 - \theta_1$
$AX = 10$	θ_2	$1 - \theta_2$
$AX = 11$	θ_3	$1 - \theta_3$

$\Leftrightarrow \{ \{1, 2\}, \{3\}, \{4\} \}$

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

L is max. likelihood, R number of parts, both w.r.t. LABELS.

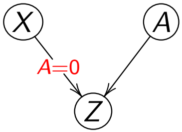
BIC for LDAGs

- Maximizing BIC [Chickering '97]:

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

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

- LABELS imply a partition of rows:



$\Leftrightarrow$

$P(Z\|A,X)$	$Z=0$	$Z=1$
$AX=00$	θ_1	$1-\theta_1$
$AX=01$	θ_1	$1-\theta_1$
$AX=10$	θ_2	$1-\theta_2$
$AX=11$	θ_3	$1-\theta_3$

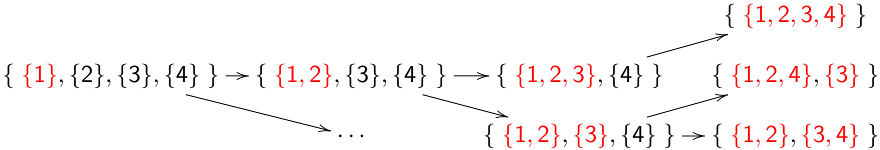
$\Leftrightarrow \{\ \{1,2\}, \{3\}, \{4\}\ \}$

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

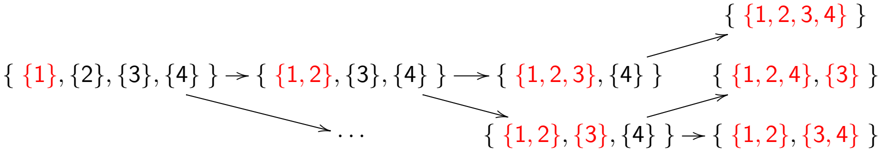
L is max. likelihood, R number of parts, both w.r.t. LABELS.

- For 4 binary parents, 27 million different label structures.

Branching . . .

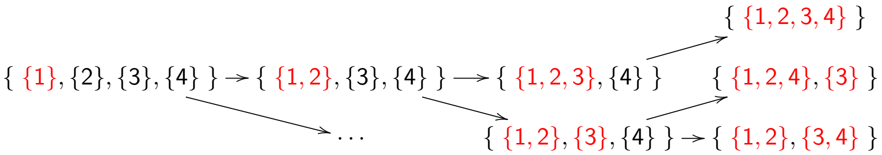


- Search over partitions of rows from complex towards simpler.



- 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

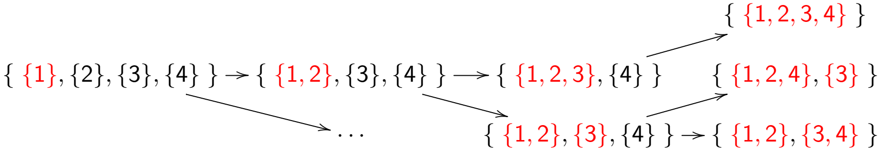


- **Upper bound** for partitions further in the branch:

$$\mathrm{L} - f \cdot \log N/2$$

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

. . . and Bounding



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

$$\mathrm{L} - f \cdot \log N/2$$

  Here L is the current likelihood, f is the number of fixed parts.
- **Initial best:** best solution to the subsets of parents.

. . . and Bounding

$\{ \{1\}, \{2\}, \{3\}, \{4\} \} \to \{ \{1,2\}, \{3\}, \{4\} \} \longrightarrow \{ \{1,2,3\}, \{4\} \} \to \{ \{1,2,3,4\} \}$

$\{ \{1\}, \{2\}, \{3\}, \{4\} \} \to \dots$

$\{ \{1,2\}, \{3\}, \{4\} \} \to \{ \{1,2\}, \{3\}, \{4\} \} \to \{ \{1,2,4\}, \{3\} \}$

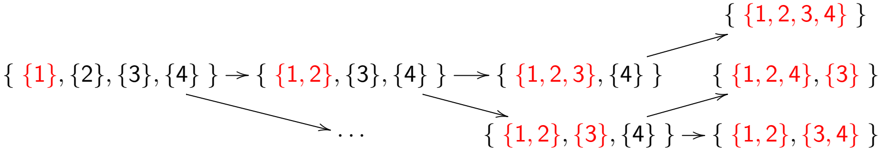
$\{ \{1,2\}, \{3\}, \{4\} \} \to \{ \{1,2\}, \{3,4\} \}$

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

$$\mathrm{L} - f \cdot \log N/2$$

  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.

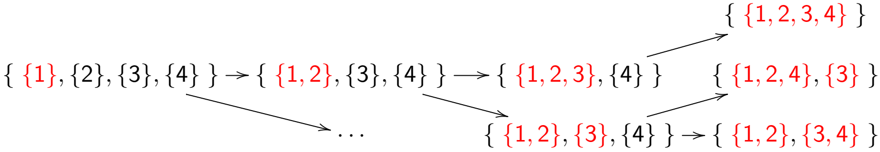
. . . and Bounding



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

$$\mathrm{L} - f \cdot \log N/2$$

  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.



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

$$\mathrm{L} - f \cdot \log N/2$$

  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:

$P(Z\|X)$	$Z=0$	$Z=1$
$X=0$	0.4	0.6
$X=1$	0.6	0.4

$\Rightarrow$

$P(Z\|X,Y)$	$Z=0$	$Z=1$
$XY=00$	0.4	0.6
$XY=01$	0.4	0.6
$XY=10$	0.4	0.6
$XY=11$	0.7	0.3

Remedies for Overfitting

- An extra edge does not always increase the BIC penalty:

$P(Z\|X)$	$Z=0$	$Z=1$
$X=0$	0.4	0.6
$X=1$	0.6	0.4

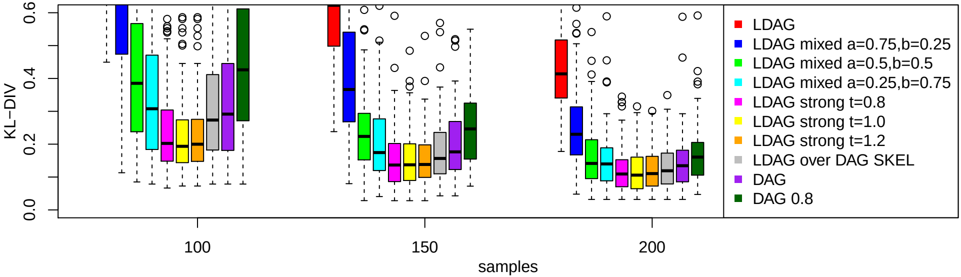
$\Rightarrow$

$P(Z\|X,Y)$	$Z=0$	$Z=1$
$XY=00$	0.4	0.6
$XY=01$	0.4	0.6
$XY=10$	0.4	0.6
$XY=11$	0.7	0.3

- **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

  $a \cdot$ LDAG-based BIC $+\, b \cdot$ DAG-based BIC.
- **LDAG over Optimal DAG Skeleton** Only orient with the LDAG-based BIC score.

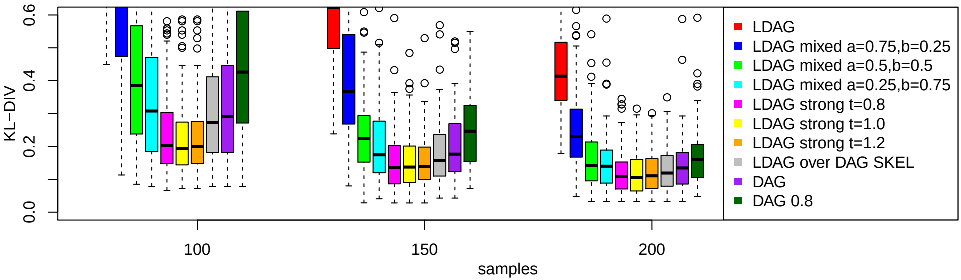
10-node binary LDAGs, 0.5 label probability. At most 3 parents.



- LDAG-based BIC overfits considerably (red).

Simulations: Probabilistic Model Accuracy

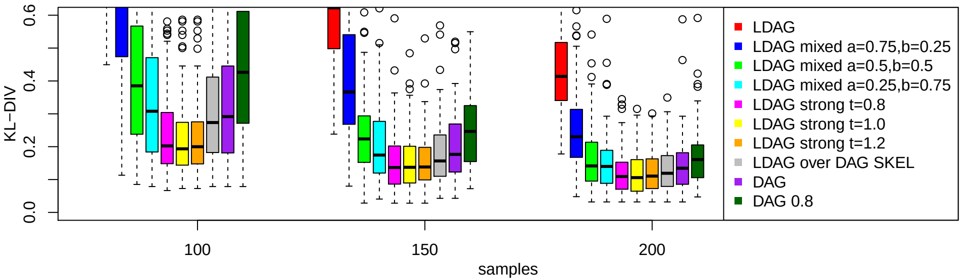
10-node binary LDAGs, 0.5 label probability. At most 3 parents.



- LDAG-based BIC overfits considerably (red).
- With strong score pruning LDAG is better than a DAG (yellow vs. purple).

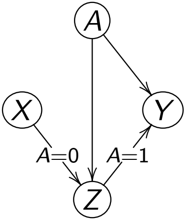
Simulations: Probabilistic Model Accuracy

10-node binary LDAGs, 0.5 label probability. At most 3 parents.



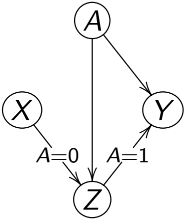
- LDAG-based BIC overfits considerably (red).
- With strong score pruning LDAG is better than a DAG (yellow vs. purple).
- With more samples DAGs catch up but still keep CSIs hidden.

Conclusion



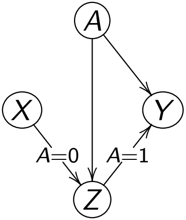
- Structure learning for labeled DAGs.

Conclusion



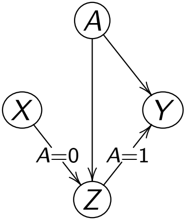
- Structure learning for labeled DAGs.
- Principled orientation of causal edges using CSIs with LPC:
  - Based on separation criteria and Markov equivalence of LDAGs.
  - More orientations more correctly than PC when CSIs present.

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.
- 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.

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.
- 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!