Learning Directed Acyclic Graphs using Integer Linear Programming

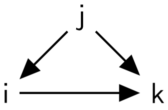
James Cussens, University of Bristol

EURO2024, 2024-07-01

Learning directed acyclic graphs (DAGs)

X_i	X_j	X_k
1.2	0.3	-4.2
-5.3	8.2	1.7
...	...	...
2.2	4.5	-9.1

$$\Longrightarrow$$



- Motivation:
  - Uncovering conditional independence relations
  - Causal Discovery

Encoding directed graphs as real vectors

- The key to the integer linear programming (ILP) approach to learning DAGs is to view them as points in $\mathbb{R}^n$.
- We do this via *family variables*.
- This digraph: $\begin{tikzcd} & j \arrow[dl] \arrow[dr] & \\ i \arrow[rr] & & k \end{tikzcd}$ is this point in $\mathbb{R}^{12}$:

$i \leftarrow \{\}$	$i \leftarrow \{j\}$	$i \leftarrow \{k\}$	$i \leftarrow \{j,k\}$
0	1	0	0

$j \leftarrow \{\}$	$j \leftarrow \{i\}$	$j \leftarrow \{k\}$	$j \leftarrow \{i,k\}$
1	0	0	0

$k \leftarrow \{\}$	$k \leftarrow \{i\}$	$k \leftarrow \{j\}$	$k \leftarrow \{i,j\}$
0	0	0	1

Encoding directed graphs as real vectors

- The key to the integer linear programming (ILP) approach to learning DAGs is to view them as points in $\mathbb{R}^n$.
- We do this via *family variables*.
- This digraph: $\begin{tikzcd} & j \arrow[dl] \arrow[dr] & \\ i \arrow[rr] & & k \end{tikzcd}$ is this point in $\mathbb{R}^{12}$:

$i \leftarrow \{\}$	$i \leftarrow \{j\}$	$i \leftarrow \{k\}$	$i \leftarrow \{j,k\}$
0	1	0	0

$j \leftarrow \{\}$	$j \leftarrow \{i\}$	$j \leftarrow \{k\}$	$j \leftarrow \{i,k\}$
1	0	0	0

$k \leftarrow \{\}$	$k \leftarrow \{i\}$	$k \leftarrow \{j\}$	$k \leftarrow \{i,j\}$
0	0	0	1

A linear objective

Many objective functions for DAG learning (e.g. BDe, BIC) are a data-determined linear function of the family variables.

$i \leftarrow \{\}$	$i \leftarrow \{j\}$	$i \leftarrow \{k\}$	$i \leftarrow \{j,k\}$
-700.1	-670.3	-630.5	-614.0

$j \leftarrow \{\}$	$j \leftarrow \{i\}$	$j \leftarrow \{k\}$	$j \leftarrow \{i,k\}$
-90.2	-40.3	-30.1	-4.2

$k \leftarrow \{\}$	$k \leftarrow \{i\}$	$k \leftarrow \{j\}$	$k \leftarrow \{i,j\}$
-20.7	-50.8	-40.9	-90.3

Score is:

A linear objective

Many objective functions for DAG learning (e.g. BDe, BIC) are a data-determined linear function of the family variables.

$i \leftarrow \{\}$	$i \leftarrow \{j\}$	$i \leftarrow \{k\}$	$i \leftarrow \{j,k\}$
-700.1	-670.3	-630.5	-614.0
0	1	0	0
$j \leftarrow \{\}$	$j \leftarrow \{i\}$	$j \leftarrow \{k\}$	$j \leftarrow \{i,k\}$
-90.2	-40.3	-30.1	-4.2
1	0	0	0
$k \leftarrow \{\}$	$k \leftarrow \{i\}$	$k \leftarrow \{j\}$	$k \leftarrow \{i,j\}$
-20.7	-50.8	-40.9	-90.3
0	0	0	1

Score is: -670.3 - 90.2 - 90.3 = -850.8

A linear objective

Many objective functions for DAG learning (e.g. BDe, BIC) are a data-determined linear function of the family variables.

$i \leftarrow \{\}$	$i \leftarrow \{j\}$	$i \leftarrow \{k\}$	$i \leftarrow \{j,k\}$
-700.1	-670.3	-630.5	-614.0
1	0	0	0
$j \leftarrow \{\}$	$j \leftarrow \{i\}$	$j \leftarrow \{k\}$	$j \leftarrow \{i,k\}$
-90.2	-40.3	-30.1	-4.2
1	0	0	0
$k \leftarrow \{\}$	$k \leftarrow \{i\}$	$k \leftarrow \{j\}$	$k \leftarrow \{i,j\}$
-20.7	-50.8	-40.9	-90.3
1	0	0	0

Score is: -700.1 - 90.2 - 20.7 = -811.0

ILP formulation for DAG learning

$$\textsc{Minimise} \quad \sum_{\substack{i \in P \\ J \subseteq P \setminus \{i\}}} c_{i \leftarrow J} x_{i \leftarrow J}$$

SUBJECT TO:

$$\sum_{J \subseteq P \setminus \{i\}} x_{i \leftarrow J} = 1 \qquad i \in P$$

$$\sum_{i \in C} \sum_{\substack{J \subseteq P \setminus \{i\} \\ J \cap C \neq \emptyset}} x_{i \leftarrow J} \leq |C| - 1 \qquad C \subseteq P, |C| \geq 2$$

$$x_{i \leftarrow J} \in \{0, 1\}, i \in P, J \subseteq P \setminus \{i\}$$

Cluster constraints

These *cluster constraints* [5] enforce acyclicity:

$$\sum_{i \in C} \sum_{\substack{J \subseteq P \setminus \{i\} \\ J \cap C \neq \emptyset}} x_{i \leftarrow J} \leq |C| - 1 \qquad C \subseteq P, |C| \geq 2$$

Here's an example where $P = \{1,2,3,4\}$ and $C = \{1,2,3\}$:

$$\begin{aligned} & x_{1\leftarrow 2} + x_{1\leftarrow 3} + x_{1\leftarrow 2,3} + x_{1\leftarrow 2,4} + x_{1\leftarrow 3,4} + x_{1\leftarrow 2,3,4} \\ + & x_{2\leftarrow 1} + x_{2\leftarrow 3} + x_{2\leftarrow 1,3} + x_{2\leftarrow 1,4} + x_{2\leftarrow 3,4} + x_{2\leftarrow 1,3,4} \\ + & x_{3\leftarrow 1} + x_{3\leftarrow 2} + x_{3\leftarrow 1,2} + x_{3\leftarrow 1,4} + x_{3\leftarrow 2,4} + x_{3\leftarrow 1,2,4} \leq 2 \end{aligned}$$

- Each cluster constraint is facet-defining.
- They are a special case of facet-defining inequalities determined by connected matroids [6].

Solving strategy

- The ILP formulation has exponentially many variables and exponentially many constraints.
- Constraints are added as cutting planes.
- Variables are added by a pricing algorithm (unless the data allows us to add in all not-fixed-at-zero variables at the start).
- Additional variables representing directed edges and ancestor relations are added to allow propagations.
- Implemented in the GOBNILP system [3, 4] which uses the SCIP library [1].

Lessons learned

- Use the fastest LP solver you can.
- One needs to trade-off the time taken to find cuts (even facet-defining ones) with the benefits of adding them.
- Most metrics measuring the success of DAG learning don't care about certificates of optimality—so adapt a suitable solving emphasis.
- Pricing is harder than cutting since we need to (re-)inspect the data.
- Carefully interleaving pricing and cutting brings benefits.

Learning from data with latent variables

$$
\begin{tikzcd}
X_1 \arrow[d] & & X_4 \arrow[d] \\
X_2 & L \arrow[l] \arrow[r] & X_3
\end{tikzcd}
$$

Two optimal DAGs learned from 100,000 datapoints (L removed) simulated from the graph above are:

$$
\begin{tikzcd}
X_1 \arrow[d] \arrow[dr] & X_4 \arrow[d] \\
X_2 \arrow[r] & X_3
\end{tikzcd}
\qquad
\begin{tikzcd}
X_1 \arrow[d] & X_4 \arrow[dl] \\
X_2 & X_3 \arrow[l] \arrow[u]
\end{tikzcd}
$$

- The true latent variable DAG can be recovered as the 'intersection' of the two optimal no-latents DAGs.
- Only worked because we had lots of data and few variables.
- Extend to less ideal learning situations?

Learning latent model directly

- Chen, Dash, and Gao [2] used an ILP approach to directly learn latent variable models: *ancestral acyclic directed mixed graphs (ADMGs)*.
- They generalise the approach presented here: instead of (binary) family variables they have (binary) variables representing *districts*.
- There are very many canidate districts and Chen, Dash, and Gao do not use pricing, which limits the approach to fairly small examples.
- Research on extending this approach is definitely worthwhile.

[1] Suresh Bolusani et al. *The SCIP Optimization Suite 9.0*. Technical Report. Optimization Online, Feb. 2024. URL: https://optimization-online.org/2024/02/the-scip-optimization-suite-9-0/.

[2] Rui Chen, Sanjeeb Dash, and Tian Gao. "Integer Programming for Causal Structure Learning in the Presence of Latent Variables". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 1550–1560. URL: https://proceedings.mlr.press/v139/chen21c.html.

[3] James Cussens. "Bayesian network learning with cutting planes". In: *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2011, pp. 153–160.

[4] James Cussens. "GOBNILP: Learning Bayesian network structure with integer programming". In: *Proceedings of the*

10th International Conference on Probabilistic Graphical Models. Ed. by Manfred Jaeger and Thomas Dyhre Nielsen. Vol. 138. Proceedings of Machine Learning Research. PMLR, 23–25 Sep 2020, pp. 605–608. URL: `https://proceedings.mlr.press/v138/cussens20a.html`.

[5] Tommi Jaakkola et al. "Learning Bayesian Network Structure using LP Relaxations". In: *Proceedings of 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*. Vol. 9. Journal of Machine Learning Research Workshop and Conference Proceedings. 2010, pp. 358–365.

[6] Milan Studený. "How matroids occur in the context of learning Bayesian network structure". In: *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI 2015)*. Ed. by Marina Meila and Tom Heskes. AUAI Press, 2015, pp. 832–841.