Labeled Directed Acyclic Graphs

Jukka Corander
Department of Mathematics and statistics
University of Helsinki, Finland
Joint work with:
Johan Pensar, Henrik Nyman and Timo Koski
Directed acyclic graphs (DAGs) have an established position as a representation of dependencies in multivariate systems.

Conditional independencies existing within a set of variables can be efficiently communicated through a DAG.

However, DAGs may still leave a substantial number of local, or context-specific independencies undiscovered.

Several approaches to generalization of DAGs have been proposed: Bayesian multinets (Geiger & Heckerman), CPT-trees (Nir Friedman et al.), Contextual weak independence (Wong & Butz), Labeled DAGs (Pensar et al.), ...
Definition

Notations for directed acyclic graphs (DAGs) to be used in statistical sense.

A DAG is $G = (V, E)$ where $V = \{1, \ldots, d\}$ is the set of nodes (stochastic variables) and $E \subseteq V \times V$ is the set of edges such that if $(i, j) \in E$ then the graph contains a directed edge from node $i$ to $j$. The set of parents of $j$ is denoted by $\Pi_j = \{i \in V : (i, j) \in E\}$. Missing edges in $G$ will be used to represent statements of conditional independences.
Definition

Conditional Independence (CI)
Let \( X = \{X_1, \ldots, X_d\} \) be a set of stochastic variables where \( V = \{1, \ldots d\} \) and let \( A, B, S \) be three disjoint subsets of \( V \). \( X_A \) is conditionally independent of \( X_B \) given \( X_S \) if

\[
p(x_A \mid x_B, x_S) = p(x_A \mid x_S)
\]

holds for all \( (x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S \) whenever \( p(x_B, x_S) > 0 \). This will be denoted by

\[
X_A \perp X_B \mid X_S.
\]
DAGs encode factorization of the joint distribution of nodes

The directed local Markov property implies that each variable $X_j$ is conditionally independent of its non-descendants given its parental variables $X_{\Pi_j}$. This leads to an explicit factorization of the joint distribution into lower order distributions,

$$p(X_1, \ldots, X_d) = \prod_{j=1}^{d} p(X_j \mid X_{\Pi_j}),$$

(1)

where the factors are CPDs associated with each node.
DAGs & CPTs

- DAG: Nodes 2, 3, 1, 4 with directed edges from 2 to 1, 3 to 1, and 1 to 4.
- CPT: Table showing conditional probabilities for $X_1 = 1$ given $X_{\Pi_j}$:

<table>
<thead>
<tr>
<th>$X_{\Pi_j}$</th>
<th>$p(X_1 = 1 \mid X_{\Pi_j})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0, 0)</td>
<td>$p_1$</td>
</tr>
<tr>
<td>(0, 0, 1)</td>
<td>$p_2$</td>
</tr>
<tr>
<td>(0, 1, 0)</td>
<td>$p_3$</td>
</tr>
<tr>
<td>(0, 1, 1)</td>
<td>$p_4$</td>
</tr>
<tr>
<td>(1, 0, 0)</td>
<td>$p_5$</td>
</tr>
<tr>
<td>(1, 0, 1)</td>
<td>$p_6$</td>
</tr>
<tr>
<td>(1, 1, 0)</td>
<td>$p_7$</td>
</tr>
<tr>
<td>(1, 1, 1)</td>
<td>$p_8$</td>
</tr>
</tbody>
</table>
DAGs cannot tell all tales...

From Geiger & Heckerman (1996):

- A guard of a secured building expects three types of persons \( h \) to approach the building's entrance: workers in the buildings, approved visitors, and spies.
- As a person approaches the building, the guard can note gender \( g \) and whether or not the person wears a badge \( b \).
- Spies are mostly men. Spies always wear badges in an attempt to fool the guard.
- Visitors don't wear badges because they don't have one.
- Female workers tend to wear badges more often than do male workers.
- The task of the guard is to identify the type of person approaching the building.
DAGs cannot tell all tales...
Context-Specific Independence (CSI)

Let $X = \{X_1, \ldots, X_d\}$ be a set of stochastic variables where $V = \{1, \ldots, d\}$ and let $A, B, C, S$ be four disjoint subsets of $V$. $X_A$ is contextually independent of $X_B$ given $X_S$ and the context $X_C = x_C$ if

$$p(x_A \mid x_B, x_C, x_S) = p(x_A \mid x_C, x_S),$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_C, x_S) > 0$. This will be denoted by

$$X_A \perp X_B \mid x_C, X_S.$$
**Definition**

**Labeled Directed Acyclic Graph (LDAG)**

Let $G = (V, E)$ be a DAG over the stochastic variables $\{X_1, \ldots, X_d\}$. For all $(i, j) \in E$, let $L_{(i,j)} = \Pi_j \setminus \{i\}$. A label on an edge $(i, j) \in E$ is defined as the set

$$\mathcal{L}_{(i,j)} = \left\{ x_{L_{(i,j)}} \in \mathcal{X}_{L_{(i,j)}} : X_j \perp X_i \mid x_{L_{(i,j)}} \right\}.$$

An LDAG is a DAG to which the label set $\mathcal{L}_E = \{\mathcal{L}_{(i,j)} : \mathcal{L}_{(i,j)} \neq \emptyset\}_{(i,j) \in E}$ has been added, it is denoted by $G_L = (V, E, \mathcal{L}_E)$.
LDAG example

\[
\mathcal{L}_{(2,1)} = (0, 1) \Rightarrow X_1 \perp X_2 \mid (X_3, X_4) = (0, 1)
\]

\[
\mathcal{L}_{(4,1)} = X_2 \times \{1\} \Rightarrow X_1 \perp X_4 \mid x_2 \in X_2, X_3 = 1
\]

\[
\Leftrightarrow X_1 \perp X_4 \mid X_2, X_3 = 1
\]

\[\begin{array}{c|c|c|c}
\hline
& x_{\mathcal{H}_1} & p(X_1 \mid x_{\mathcal{H}_1}) \\
\hline
X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0 & & p_1 \\
X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1 & & p_3 \\
X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 0 & & p_4 \\
X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 1 & & p_4 \\
X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0 & & p_2 \\
X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1 & & p_3 \\
X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 0 & & p_5 \\
X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1 & & p_5 \\
\hline
\end{array}\]
The previous LDAG is also representable as a CPT-tree (Boutilier et al.)

\[
\begin{array}{c|c|c|c|c}
\hline
x_{II_1} & p(X_1 | x_{II_1}) \\
\hline
X_3 = 0 \land X_4 = 0 \land X_2 = 0 & p_1 \\
X_3 = 0 \land X_4 = 0 \land X_2 = 1 & p_2 \\
X_3 = 0 \land X_4 = 1 & p_3 \\
X_3 = 1 \land X_2 = 0 & p_4 \\
X_3 = 1 \land X_2 = 1 & p_5 \\
\hline
\end{array}
\]

\[X_3 = 0 \land X_4 = 1 \Rightarrow X_1 \perp X_2 \mid (X_3, X_4) = (0, 1)\]
\[X_3 = 1 \land X_2 = 0 \Rightarrow X_1 \perp X_4 \mid (X_2, X_3) = (0, 1)\]
\[X_3 = 1 \land X_2 = 1 \Rightarrow X_1 \perp X_4 \mid (X_2, X_3) = (1, 1)\]
CPT-tree is an LDAG but CPT-trees do not cover all LDAGs
**Definition**

**Maximal LDAG**
An LDAG $G_L = (V, E, \mathcal{L}_E)$ is called maximal if there exists no configuration $x_{L(i,j)}$ that can be added to the label $\mathcal{L}_{(i,j)}$ without inducing an additional local CSI. If we add a configuration to a label in a maximal LDAG, it must result in an additional restriction in the form of a local CSI. This will in turn result in a reduction of the associated minimal reduced CPT.

**Definition**

**Regular maximal LDAG**
A maximal LDAG $G_L = (V, E, \mathcal{L}_E)$ is regular if $\mathcal{L}_{(i,j)}$ is a strict subset of $\mathcal{X}_{L(i,j)}$ for every label in $G_L$. 
LDAGs revisited...

**Theorem**

Consequence of maximality, essential for interpretability of LDAGs
Let $G_L = (V, E, \mathcal{L}_E)$ and $G_L^* = (V, E, \mathcal{L}_E^*)$ be two maximal LDAGs with the same underlying DAG $G = (V, E)$. Then $G_L$ and $G_L^*$ represent equivalent dependence structures if and only if $\mathcal{L}_E = \mathcal{L}_E^*$, i.e. $G_L = G_L^*$.

**Theorem**

In a regular maximal LDAG, a label $\mathcal{L}_{(i,j)}$ cannot induce an independence assertion of the form $X_j \perp X_i \mid x_{L(i,j)}$ for all $x_{L(i,j)} \in \mathcal{X}_{L(i,j)}$, i.e. $X_j \perp X_i \mid X_{L(i,j)}$.
Illustration of previous results

Label (1,1) can be added

Here the edge vanishes
Central concepts for Markov properties in LDAGs

Definition

Satisfied label
Let $G_L = (V, E, \mathcal{L}_E)$ be an LDAG and $X_C = x_C$ a context where $C \subseteq V$. In the context $X_C = x_C$, a label $L_{(i,j)} \in \mathcal{L}_E$ is satisfied if $L_{(i,j)} \cap C \neq \emptyset$ and $\{x_{L_{(i,j)} \cap C}\} \times X_{L_{(i,j)} \setminus C} \subseteq L_{(i,j)}$.

Definition

Context-specific LDAG
Let $G_L = (V, E, \mathcal{L}_E)$ be an LDAG. For the context $X_C = x_C$, where $C \subseteq V$, the context-specific LDAG is denoted by $G_L(x_C) = (V, E \setminus E', \mathcal{L}_{E \setminus E'})$ where $E' = \{(i, j) \in E : L_{(i,j)} \text{ is satisfied}\}$. The underlying DAG of the context-specific LDAG is denoted by $G(x_C) = (V, E \setminus E')$. 
Generalization of d-separation by Boutilier et al. – but it is NOT complete!

**Definition**

*CSI-separation in LDAGs*

Let $G_L = (V, E, L_E)$ be an LDAG and let $A, B, S, C$ be four disjoint subsets of $V$. $X_A$ is CSI-separated from $X_B$ by $X_S$ in the context $X_C = x_C$ in $G_L$, denoted by

$$X_A \perp X_B \, |_{G_L} \, x_C, X_S,$$

if $X_A$ is $d$-separated from $X_B$ by $X_{S\cup C}$ in $G(x_C)$. 
Central results for LDAGs

**Definition**

*CSI-equivalence for LDAGs*

All DAGs within the same Markov equivalence class share the same dependence structure or $\mathcal{I}(G)$. Let $G_L = (V, E, \mathcal{L}_E)$ and $G'_L = (V, E^*, \mathcal{L}'_E)$ be two distinct regular maximal LDAGs. The LDAGs are said to be CSI-equivalent if $\mathcal{I}(G_L) = \mathcal{I}(G'_L)$. A set containing all CSI-equivalent LDAGs forms a CSI-equivalence class.
Central results for LDAGs

**Theorem**

The-skeleton-condition

Let $G_L = (V, E, \mathcal{L}_E)$ and $G^*_L = (V, E^*, \mathcal{L}_E^*)$ be two regular maximal LDAGs belonging to the same CSI-equivalence class. Their underlying DAGs $G = (V, E)$ and $G^* = (V, E^*)$ must then have the same skeleton.
The equivalence theorem
Let $G_L = (V, E, \mathcal{L}_E)$ and $G_L^* = (V, E^*, \mathcal{L}_E^*)$ be two maximal regular LDAGs for which there exists distributions $P$ and $P^*$ such that $\mathcal{I}(G_L) = \mathcal{I}(P)$ and $\mathcal{I}(G_L^*) = \mathcal{I}(P^*)$. $G_L$ and $G_L^*$ are CSI-equivalent if and only if their corresponding context-specific LDAGs $G_L(x_V) = G(x_V)$ and $G_L^*(x_V) = G^*(x_V)$ are Markov equivalent for all $x_V \in \mathcal{X}_V$. 
Reality comes finally into play...

Definition

*Data and model parametrization*

Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ denote a set of training data consisting of $n$ observations $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ of the variables $\{X_1, \ldots, X_d\}$ such that $\mathbf{x}_i \in \mathcal{X}$. $\Theta_{G_L}$ is the parameter space induced by an LDAG and an instance $\theta \in \Theta_{G_L}$ corresponds to a specific joint distribution that factorizes according to the LDAG $G_L$.

The CSI-consistent partition of the outcome space $\mathcal{X}_{\Pi_j}$ is denoted by $S_{\Pi_j} = \{S_{j1}, \ldots, S_{jk_j}\}$ where $k_j = |S_{\Pi_j}|$ is the number of outcome classes. Let $r_j = |\mathcal{X}_j|$ and $q_j = |\mathcal{X}_{\Pi_j}|$ denote the cardinality of the outcome space of variable $X_j$ and its parents $X_{\Pi_j}$, respectively.

Finally, $n(\mathbf{x}_{ij} \times S_{jl})$ denotes the total count of the configurations $\{x_{ij}\} \times S_{jl}$ in $\mathbf{X}$. 

Aalto University
Bayesian learning (non-reversible MCMC+HC used for optimization)

**Definition**

*LDAG posterior and our estimate*

\[ p(G_L \mid X) = \frac{p(X \mid G_L) \cdot p(G_L)}{\sum_{G_L \in \mathcal{G}_L} p(X \mid G_L) \cdot p(G_L)}. \]

\[ p(G_L) \propto \kappa^{\text{dim}(\Theta_{G_L})} = \prod_{j=1}^{d} \kappa^{\text{dim}(\Theta_{G_L}(j))} \]

\[ \arg \max_{G_L \in \mathcal{G}_L} p(X \mid G_L) \cdot p(G_L). \]

**Lemma**

*Generalization of the Cooper-Herskovitz Lemma*

\[ p(X \mid G_L) = \int_{\theta \in \Theta_{G_L}} p(X \mid G_L, \theta) \cdot f(\theta \mid G_L) \, d\theta \]

\[ p(X \mid G_L) = \prod_{j=1}^{d} \prod_{i=1}^{k_j} \frac{\Gamma\left(\sum_{l=1}^{r_j} \alpha_{ijl}\right)}{\Gamma\left(n(S_{ji}) + \sum_{l=1}^{r_j} \alpha_{ijl}\right)} \prod_{i=1}^{r_j} \frac{\Gamma\left(n(x_{ji} \times S_{ji}) + \alpha_{ijl}\right)}{\Gamma(\alpha_{ijl})} \]
Some empirical results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>Smoking</td>
<td>No := 0, Yes := 1</td>
</tr>
<tr>
<td>$X_2$</td>
<td>Strenuous mental work</td>
<td>No := 0, Yes := 1</td>
</tr>
<tr>
<td>$X_3$</td>
<td>Strenuous physical work</td>
<td>No := 0, Yes := 1</td>
</tr>
<tr>
<td>$X_4$</td>
<td>Systolic blood pressure</td>
<td>$&lt; 140 := 0, &gt; 140 := 1$</td>
</tr>
<tr>
<td>$X_5$</td>
<td>Ratio of $\beta$ and $\alpha$ lipoproteins</td>
<td>$&lt; 3 := 0, &gt; 3 := 1$</td>
</tr>
<tr>
<td>$X_6$</td>
<td>Family anamnesis of coronary heart disease</td>
<td>No := 0, Yes := 1</td>
</tr>
</tbody>
</table>

Optimal LDAG for coronary heart disease data.
Some empirical results
DAGs keep evolving...

(minor revision submitted to Data Mining and Knowledge Discovery)