Modeling uncertain interventions

Kevin Murphy
U. British Columbia

Joint work with
David Duvenaud, Guillaume Alain, Daniel Eaton
Outline

• Reducing causality to decision theory
• Learning DAGs with “fat hands”
• Beyond DAGs
2 types of causality

• Phil Dawid distinguishes 2 types of causality

• Effects of Causes
  – e.g., if I take an aspirin now, will that cause my headache to go away?

• Causes of Effects
  – e.g., my headache has gone; would it be gone if I had not taken the aspirin?

- “Causal inference without counterfactuals”, JASA 2000
Causality -> decision theory

• Most applications of causal reasoning are concerned with Effects of Causes. This can be modeled using standard decision theory.

• Reasoning about Causes of Effects requires counterfactuals, which are fundamentally unidentifiable, hence dangerous.

• We shall focus of Effects of Causes (Pearl 2000, ch 1-6).
Intervention DAGs

- Each intervention/action $A_j$ node determines if $X_j$ is sampled from its normal or ‘mutated’ mechanism.
- Perfect intervention means cutting incoming arcs: $p(X_2 | X_1, A_2=1, \theta_2) = \delta(X_2 - \theta_2^1)$

![Diagram of a DAG with nodes $A_1$, $X_1$, $A_2$, and $X_2$, and edges connecting $A_1$ to $X_1$, $A_2$ to $X_2$, and $X_1$ to $X_2$. The node $X_1$ has parameters $\theta_1^0$, $\sigma_1$, $\theta_1^1$, and the node $X_2$ has parameters $\theta_2^0$, $\sigma_2$, $\theta_2^1$.]


Observing vs doing

• I-DAGs make the do-operator and edge-cutting unnecessary

\[ p(X_1 | X_2 = x_2) = p(X_1 | X_2 = x_2, A_1 = 0, A_2 = 0) \]
\[ p(X_1 | \text{do}(X_2 = x_2)) = p(X_1 | X_2 = x_2, A_1 = 0, A_2 = 1) \]
Distinguishing causally different DAGS

- I-DAGs can resolve Markov equivalence
Back-door criterion

- D-separation in I-DAG can be used to derive all of Pearl’s results (in ch1-6) and more

\[ A_t \rightarrow T \leftarrow C \]

\[ C \perp A_t \]

\[ R \perp A_t | C, T \]

\[ p(r|A_t = t) = \sum_c p(r|A_t = t, c)p(c|A_t = t) \]

\[ = \sum_c p(r|A_t = 0, T = t, c)p(c|A_t = 0) \]

Dawid, Intl Stat Review 2002
Structure learning

• Posterior over graphs given interventional and observational data:

\[ p(G|X, A) \propto p(G)p(X|G, A) \]

\[ = p(G) \prod_{j=1}^{d} \int_{i:A_{ij}=0} \prod_{i} p(X_{ij}|X_i, G_j, \theta_j) d\theta_j \]

• We just modify the marginal likelihood (or BIC) criterion to exclude training cases where node was set by intervention

Cooper & Yoo, UAI’99
"Causal Protein-Signaling Networks derived from multiparameter Single-Cell Data", Sachs, Perez, Pe’er, Lauffenberger, Nolan, Science 2005
Aside on algorithms

• Sachs et al. used simulated annealing
• Ellis & Wong used equi-energy sampling
• Eaton & Murphy used dynamic programming (Koivisto) to compute the exact posterior mode and exact edge marginals $p(G_{ij}=1|X,A)$.
• Can use DP as proposal for MH.

- Byron Ellis and Wing Wong, “Learning causal Bayesian networks from experimental data”, JASA 2008
- Daniel Eaton and Kevin Murphy, “Exact Bayesian structure learning from uncertain interventions”, AI/Stats 2006
- “Advances in exact Bayesian structure discovery in Bayesian networks”, M. Koivisto, UAI 2006
Error vs compute time (5 nodes)

$$\sum_{i,j} |p(G_{ij} = 1|D) - \hat{p}_t(G_{ij} = 1|D)|$$

- Eaton & Murphy,” Bayesian structure learning using dynamic programming and MCMC”, UAI 2007
Outline

• Reducing causality to decision theory
• Learning DAGs with “fat hands”
• Beyond DAGs
T-cell interventions

Sachs et al, Science ‘05
Intervening on hidden variables
Actions appear as “fat hands”

MAP DAG computed exactly by DP from large training set
Thin vs fat hands
Thin vs fat in T-cell example

“Ground truth” DAG
DAG learned with perfect intervention assumption is quite similar…

Learned fat-hand DAG
Samples from learned models

Training
Samples from learned models

Posterior predictive checking, without reference to “ground truth” DAG
Cross validation

- Negative log-likelihood on 10-fold CV

- Learning effects of intervention is better than assuming they are perfect.

Eaton & Murphy, AI/Stats 2006
Aside on algorithms

• The DAG is block-structured, since no X→A or A→A edges.
• Can exploit this in the DP algorithm so computation is $O(2 \cdot 2^d)$ not $O(2^{2d})$
Outline

• Reducing causality to decision theory
• Learning DAGs with “fat hands”
• Beyond DAGs
I-DAGs represent $p(x|a)$

- DAGs are a way of representing joint distributions $p(x)$ in factored form.
- I-DAGs are a way of representing conditional distributions $p(x|a)$ in factored form, assuming actions have local effects.
- This lets us fit fewer than $O(2^d)$ separate distributions, so we can pool data, and allows us to generalize to new conditioning cases.
Predicting fx of novel interventions

• Main focus of current literature: predict effects of interventions given observational data, i.e., predict \( p(x | \text{do}(x_j)) = p(x | a_j=1, a_{-j}=0) \) given samples from \( p(x | a=0) \)

• Other possible questions: predict \( p(x | a_j=1, a_k=1, a_{-jk}=0) \) given samples from \( p(x | a_j=1, a_j=0, a_{-jk}=0) \) and \( p(x | a_j=0, a_k=1, a_{-jk}=0) \)
DREAM 3 signaling response challenge

Predict value of 17 phosphoproteins and 20 cytokines at 3 time points in 2 cell types under novel combinations of stimulus/inhibitor

Dialogue on Reverse Engineering and Assessment Methods
How to Fill in a Matrix?

• Need to borrow statistical strength from your (unordered) row and column neighbours

• Almost like predicting what rating someone will give a movie...

\[
\begin{pmatrix}
3 & 2 & 4 & ? \\
8 & 5 & 7 & ? \\
9 & ? & 6 & 9
\end{pmatrix}
\]
Probabilistic Matrix Factorization

- Singular Value Decomposition with missing entries, plus some L2 regularization

\[
\begin{align*}
R_{ij} &= u_i^T v_j + \epsilon_{ij} \\
&= \sigma_u u_i \sigma_v v_j + \epsilon_{ij}
\end{align*}
\]

\[
\begin{bmatrix}
3 & 2 & 4 & ? \\
8 & 5 & 7 & ? \\
9 & ? & 6 & 9
\end{bmatrix}
\]

Salakhutdinov & Mnih, NIPS 2007
Linear regression

If $u_i, v_j$ are scalar, we can use linear regression

\[
\begin{pmatrix}
1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
u_1 \\
u_2 \\
u_3 \\
v_1 \\
v_2 \\
v_3 \\
v_4
\end{pmatrix}
=
\begin{pmatrix}
u_1 + v_1 \\
u_1 + v_2 \\
u_1 + v_3 \\
u_2 + v_1 \\
u_2 + v_2 \\
u_3 + v_3 \\
u_3 + v_1 \\
u_3 + v_3 \\
u_3 + v_4
\end{pmatrix}
=
\begin{pmatrix}
3 \\
2 \\
4 \\
8 \\
5 \\
7 \\
9 \\
6 \\
9
\end{pmatrix}
\begin{pmatrix}
3 & 2 & 4 & \_ \\
8 & 5 & 7 & \_ \\
9 & \_ & 6 & 9
\end{pmatrix}
## Linear regression for dream 3

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Inhibitor</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>MEK</td>
<td>P38</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th>Team</th>
<th>Normalized Squared Error</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBC PMF</td>
<td>1483.961</td>
<td>2.116e-024</td>
</tr>
<tr>
<td>UBC Index-Based</td>
<td>1828.389</td>
<td>5.771e-024</td>
</tr>
<tr>
<td>Team 102</td>
<td>3101.950</td>
<td>2.080e-022</td>
</tr>
<tr>
<td>Team 106</td>
<td>3309.644</td>
<td>3.682e-022</td>
</tr>
<tr>
<td>Team 302</td>
<td>11329.398</td>
<td>7.365e-014</td>
</tr>
</tbody>
</table>

• We won!
• However, the contest was already over 😒
• Also, none of the other methods used DAGs...
• How do these simple methods compare to DAG-based approaches on the T-cell data?
Modified T-cell data

We have observations of $X_{i,1:d}$ for $i=1:1000$, given $A_a=1$, $A_{-a}=0$, for $a=1:6$. From this, compute average response of each variable to each action.

$$\mu_{a,j} = \frac{1}{n_a} \sum_{i: A_a = 1} x_{i,j}$$
Predictive accuracy on modified T-cell
Predictive accuracy on modified T-cell
Predictive accuracy on modified T-cell
Predictive accuracy on modified T-cell
Summary

• Effects of Causes can be modeled using influence diagrams, which can be learned from data using standard techniques.

• Other kinds of conditional density models can also be used, and work surprisingly well.

• We need to assess performance without reference to graph structures, which, for real data, can never be observed.