Introduction
Chain Event Graphs
Model selection for CEGs
Conclusions

Chain Event Graphs for Informative Missingness

Jane Hutton, Lorna Barclay, Jim Smith

University of Warwick

MRC Conference on Biostatistics
Graphical Models

- Chain Event Graphs are a type of graphical model.
- They are derived from probability trees by merging the nodes in a tree whose associated conditional probabilities are the same.
- The use of a graph makes the statistical model more accessible by providing a visual representation of the problem.

$\Rightarrow$ We show that the CEG allows us to draw informative conclusions directly from its graph.
$\Rightarrow$ It can represent missing data structures explicitly.
$\Rightarrow$ It can provide a useful framework for defining categories of variables.
Example: Christchurch Health and Development Study

‘Effect of family and substance use on hospital admission’

- Birth cohort study following up 1265 children born in Christchurch, New Zealand
- Hospital admissions aged 21-25 years, not pregnancy related.

Effect of family type at birth, substance abuse aged 16-18 years on hospital admissions

- $X_2 = \text{Family type at birth: both parents, single, adopted}$
- $X_3 = \text{Substance abuse aged 16-18: none, user, missing}$
- $X_4 = \text{Hospital admission: no admission, at least one admission}$
CEG: Initial tree

- **Parents:**
  - both
  - single
  - adoptive

- **Drug abuse age 16-18:**
  - no
  - user
  - missing

- **Hospital admissions:**
  - none
  - some or missing

Lines have the same colour if they represent the same conditional probability.

Figure: Tree on three variables
Two situations $v, v'$ are in the same **stage** $u$ if and only if
- The topology of their florets $F(v)$ and $F(v')$ are the same
- There is a bijection between the florets such that the probabilities on corresponding edges are the same

Two situations $v, v'$ are in the same **position** $w$ if and only if
- The topology of their subtrees $T(v)$ and $T(v')$ are the same
- There is a bijection between the subtrees such that the probabilities on corresponding edges are the same
CEG: Stages and Positions

- **Stages:**

  \[ u_1 = \{ v_1 \}, \quad u_2 = \{ v_2, v_3, v_4 \}, \]
  \[ u_3 = \{ v_5, v_7 \}, \quad u_4 = \{ v_6, v_8, v_9 \}, \]
  \[ u_5 = \{ v_{10} \} \]

- **Positions:**

  \[ w_1 = \{ v_1 \}, \quad w_2 = \{ v_2, v_3 \}, \]
  \[ w_3 = \{ v_4 \}, \quad w_4 = \{ v_5, v_7 \}, \]
  \[ w_5 = \{ v_6, v_8, v_9 \}, \quad w_6 = \{ v_{10} \} \]
  \[ w_\infty = \{ l_1, l_2, l_3, \ldots, l_{10}, l_{11}, l_{12} \} \]

**Figure:** Tree on three variables

Hutton, Barclay, Smith

CEGs for missing data
Definition of a Chain Event Graph

- The set of vertices is the set of all positions of the tree $T$ and the position of all leaf nodes.
- For each position $w$ choose a single representative situation $v(w)$. We have an edge from $w$ to $w'$ for each edge from $v(w)$ to a vertex $v' \in w'$.
- If $u(w) \neq \{w\}$, there is more than one position in the stage, so we connect two positions by an undirected dotted line.

**Figure:** Tree on three variables

---

Hutton, Barclay, Smith

CEGs for missing data
Definition of a Chain Event Graph

$w_1$ = Root, $w_2$ = Both or adoptive parents, $w_3$ = Single parent, $w_4$ = Both or adoptive parents and no drug abuse, $w_5$ = (Both or adoptive parents and drug abuse) or (Single parent and no drug abuse), $w_6$ = Single parent and drug abuse, $w_\infty$ = Sink.

Figure: CEG derived from the tree, $T$
CEG: Missing data

- Tree on three variables, with missing data
- Percent no admissions to hospital
**Definition:** Let $T$ be a tree on $p$ variables with a binary outcome variable $Y_p$ represented by the leaf nodes in the tree. A CEG, $C(T)$, is an ordinal CEG with respect to $Y_p$ when the positions in each vertex subset associated with a variable $Y_i$, $V_{Y_i}$, are vertically aligned in descending order with respect to the predictive probability $P(Y_p = 0 | D, C(T))$.

Figure: Ordinal CEG when data are MAR

Hutton, Barclay, Smith

CEGs for missing data
Ordinal CEG: Missing completely at random

The positions $w_2$, $w_3$ and $w_4$ are in the same stage.

**Figure**: Ordinal CEG when data are MCAR
Ordinal CEG: Missing not at random

Figure: Ordinal CEG when data are MNAR
Score of the CEG

- Let $\Pi_u$ be the set of conditional probabilities associated with the floret $F(u)$
- $\Pi_u \sim \text{Dir}(\alpha_u)$, $\alpha_u = (\alpha_{u1}, \ldots, \alpha_{ur_u})$
- $\Pi_u|D \sim \text{Dir}(\alpha_u + N_u)$, $N_u = (N_{u1}, \ldots, N_{ur_u})$
- Simplest case: uniform prior on the root-to-leaf paths of the associated tree

The score of a CEG structure $C$ given a dataset $D$ is

$$\log L(C|D) = \sum_{u \in J(T)} \left( \log \Gamma(\alpha_u) - \log \Gamma(\alpha_u + N_u) + \sum_{k=1}^{r_u} \{ \log \Gamma(\alpha_{uk} + N_{uk}) - \log \Gamma(\alpha_{uk}) \} \right)$$

We compare two CEG structures using log Bayes factors:

$$\log L(C_1|D) - \log L(C_0|D).$$
Example: Christchurch Health and Development Study

‘Effect of family and substance use on hospital admission’

![CEG diagram]

**Figure:** CEG for example; % no reported hospital admissions
Conclusions

⇒ The CEG allows us to draw informative conclusions directly from its graph
⇒ It can represent missing data structures explicitly
⇒ It can provide a useful framework for defining categories of variables

Further work:
- Informative priors
- Include more covariates
Ordinal CEG: Conditional Missing not at random

Figure: Ordinal CEG when data are MNAR conditional on birth weight