Complex Systems Methods — 5. Statistical complexity, conditional independence and graphical models

Eckehard Olbrich

e.olbrich@gmx.de
http://personal-homepages.mis.mpg.de/olbrich/complex_systems.html

Potsdam WS 2007/08

Overview

Summary: Complexity measures for finite systems

- "World": a set V of $1 \leq N < \infty$ elements (agents, nodes) with state sets $\mathcal{X}_v$, $v \in V$.
- Given a probability vector p on $\mathcal{X}_V$ we get random variables X_V on V, X_A on $A \subseteq V$ and X_v on $v \in V$.
- Measuring statistical dependencies: Integration or Multi-information

$$I(X_V) := \sum_{v \in V} H(X_{\{v\}}) - H(X_V) = D\left(p(x_V) || \prod_{v \in V} p_v(x_{\{v\}})\right)$$

- Excess entropy

$$E(X_V) := H(X_V) - \sum_{v \in V} H(X_{\{v\}} | X_{V \setminus \{v\}})$$

Summary: Complexity measures for finite systems

- "Neural complexity" introduced by Tononi, Sporns and Edelman (1994) — TSE-complexity:

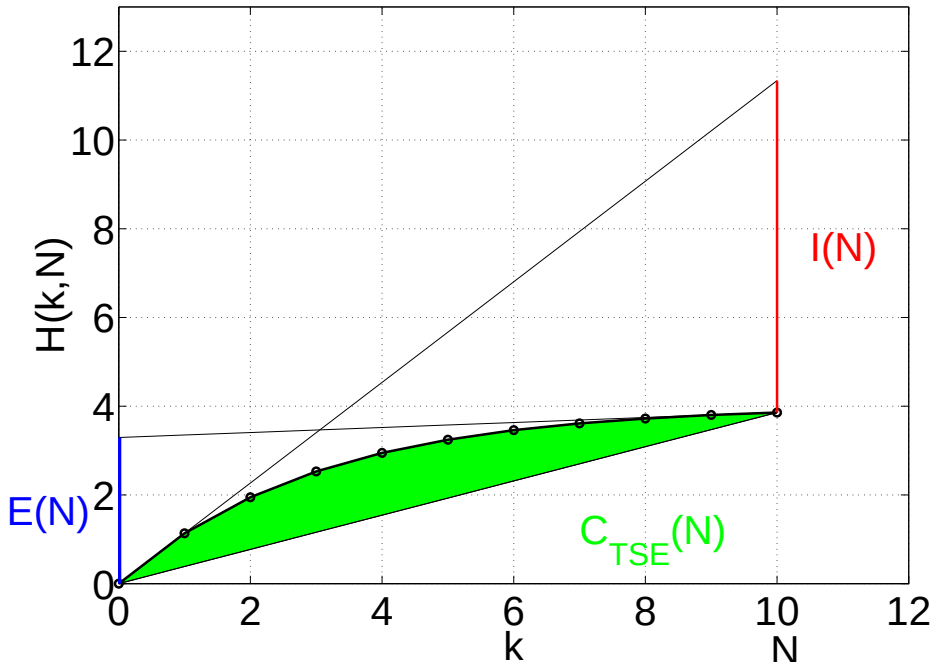
$$C_{TSE}(X_V) = \sum_{k=1}^{N} \left(H(k,N) - \frac{k}{N} H(N) \right) \quad H(k,N) = \binom{N}{k}^{-1} \sum_{\substack{Y \subseteq V \\ |Y|=k}} H(X_Y)$$

- High TSE-complexity requires low integration for small subsystems and high integration at the system level:

$$C_{TSE}(X_V) := \sum_{k=1}^{N} \left(\frac{k}{N} I(N) - I(k,N) \right)$$

$$I(k,N) = \binom{N}{k}^{-1} \sum_{\substack{Y \subseteq V \\ |Y|=k}} \left(\sum_{v \in Y} H(_{\{v\}} - H(X_Y) \right)$$

Integration, excess entropy and TSE-complexity



TSE-complexity can be expressed as the sum over the averaged excess entropy of subsystems of size k.

$$C_{TSE}(X_V) = \frac{1}{2}\sum_{k=1}^{N} E(k,N) = \frac{1}{2}\sum_{Y\subseteq V} \frac{1}{\binom{N}{|Y|}} E(X_Y)$$

Some properties of the excess entropy for finite systems

1. $E(X_V) = H(X_V) - \sum_{v \in V} H(X_{\{v\}}|X_{V\setminus\{v\}}) \leq H(X_V)$
2. The excess entropy of a system consisting of two subsystems A and B is always larger than the mutual information between these two subsystems:
$$E(X_{A\cup B}) \geq I(X_A : X_B) \; .$$
3. The excess entropy of the union of two subsystems is always larger than the excess entropy of one of the subsystems.
$$E(X_{A\cup B}) \geq E(X_A) \qquad E(X_{A\cup B}) \geq E(X_B)$$
4. The sum of the excess entropies of the subsystems can be either less or larger than the excess entropy of the whole system.
$$\begin{aligned} E(X_{A\cup B}) &= E(X_A) + E(X_B) + \sum_{v\in A} I(X_{\{v\}} : X_B|X_{A\setminus\{v\}}) + \\ &\quad + \sum_{v\in X_B} I(X_{\{v\}} : X_A|X_{B\setminus\{v\}}) - I(X_A : X_B) \; . \end{aligned}$$

Chain rules - Partioning the system

Divide the system into finer and finer partitions according to the following rule:

1. *Initialization:* Start the sequence of partitions by defining as first partition the trivial one: $\xi_1 := \{V\}$
2. *Step* $k \to k+1$*:* If all atoms of the partition ξ_k have exactly one element, then stop. Otherwise, choose one atom A_k of the partition ξ_k that has at least two elements and divide it into two non-empty and disjoint sets A_k^1 and A_k^2 with $A_k = A_k^1 \cup A_k^2$. Define the new partition ξ_{k+1} according to

$$\xi_{k+1} := \left(\xi_k \setminus \{A_k\}\right) \cup \{A_k^1, A_k^2\}$$

3. Go to the second step. This procedure generates a sequence of bipartitions $A_k = A_k^1 \cup A_k^2$

Chain rules

Integration: For all k we have the decomposition rule

$$I(X_{A_k}) = I(X_{A_k^1} : X_{A_k^2}) + I(X_{A_k^1}) + I(X_{A_k^2}) ,$$

which finally leads to the chain rule for multi-information

$$I(X_V) = \sum_{k=1}^{N-1} I(X_{A_k^1} : X_{A_k^2}) .$$

Excess entropy: We have the following decomposition:

$$E(X_V) = \sum_{k=1}^{N-1} I(X_{A_k^1} : X_{A_k^2} | X_{V \setminus (A_k^1 \cup A_k^2)}) .$$

Similar terms in both expressions, but in the first case unconditioned and in the second conditioned mutual information.

Excess entropy and conditional independence

- Time series (forecasting complexity):

$$E_N = \sum_{k=1}^{N} k \cdot \delta h_k = \sum_{k=1}^{\infty} kMI(X_0 : X_{-k}|X_{-1}, \ldots, X_{-k+1})$$

- General case: Chain rule

$$E(X_V) = \sum_{k=1}^{N-1} I(X_{A_k^1} : X_{A_k^2}|X_{V\setminus(A_k^1 \cup A_k^2)})$$

or with an (arbitrary) ordering of the nodes

$$E_N = \sum_{k=1}^{N} \sum_{j=k+1}^{N} MI(x_k; x_j|, x_{k+1}^{j-1}, x_1^{k-1}) \, .$$

- Conditional independence $\Rightarrow$ Markov property $\Rightarrow$ less terms in the sums for the excess entropy $\Rightarrow$ lower complexity
- Conditional independence simplifies statistical dependencies
- This can be visualized by graphs $\Rightarrow$ Graphical models

Conditional independence and conditional mutual information

- Conditional independence: X is conditional independent on Y given Z, written $X \perp\!\!\!\perp Y|Z$, if $p(X|Y,Z) = p(X|Z)$, i.e. Y is irrelevant for explaining X if Z is already known. $X \perp\!\!\!\perp Y|Z \Leftrightarrow MI(X:Y|Z) = 0$.
- Some properties:

(1) Symmetry	$X \perp\!\!\!\perp Y\|Z \Rightarrow Y \perp\!\!\!\perp X\|Z$
(2) Decomposition	$X \perp\!\!\!\perp YW\|Z \Rightarrow X \perp\!\!\!\perp Y\|Z$
(3) Weak union	$X \perp\!\!\!\perp YW\|Z \Rightarrow X \perp\!\!\!\perp Y\|ZW$
(4) Contraction	$X \perp\!\!\!\perp Y\|Z \;\&\; X \perp\!\!\!\perp W\|ZY \Rightarrow X \perp\!\!\!\perp YW\|Z$
(5) Intersection	$X \perp\!\!\!\perp W\|ZY \;\&\; X \perp\!\!\!\perp Y\|ZW \Rightarrow X \perp\!\!\!\perp YW\|Z$ if $p(X,Y,Z,W) > 0$

- (1)-(4) can be shown using the symmetry (1) and the chain rule (2,3,4) for the conditional mutual information.
- These properties are called the *Graphoid axioms*.

Second order conditional independence

- Correlation function $C_{XY} = E[XY]$.
- Partial correlation function $E[(X - \tilde{X}(Z))(Y - \tilde{Y}(Z))]$ with $\tilde{X}(Z)$ denoting the best linear prediction of X from Z.
- Let us denote by $X \perp\!\!\!\perp_2 Y|Z$ that the partial correlation function between X and Y given Z vanishes. Then $\perp\!\!\!\perp_2$ satisfies also the graphoid axioms.
- Also some graph properties satisfy this axioms and can therefore be used to represent conditional independence $\Rightarrow$ Graphical models
- These properties are called *Markov properties*.

Graphical Models

- **Graphical Modelling** is an area which has its roots in statistics, but which also incorporates neural networks, hidden Markov models, and many other techniques that exploit **conditional independence** properties for modelling, display, and computation.
- Using conditional independence assumptions the analysis of high-dimensional problems can be split up into small manageable pieces, introducing some kind of "modularity".
- These conditional independence structures can be represented graphically. The resulting networks are often calles *Bayesian networks*, a slightly more general term is *Probabilistic networks*.
- Graphical model represents the qualitative structure of a problem.

Example: Expert Systems

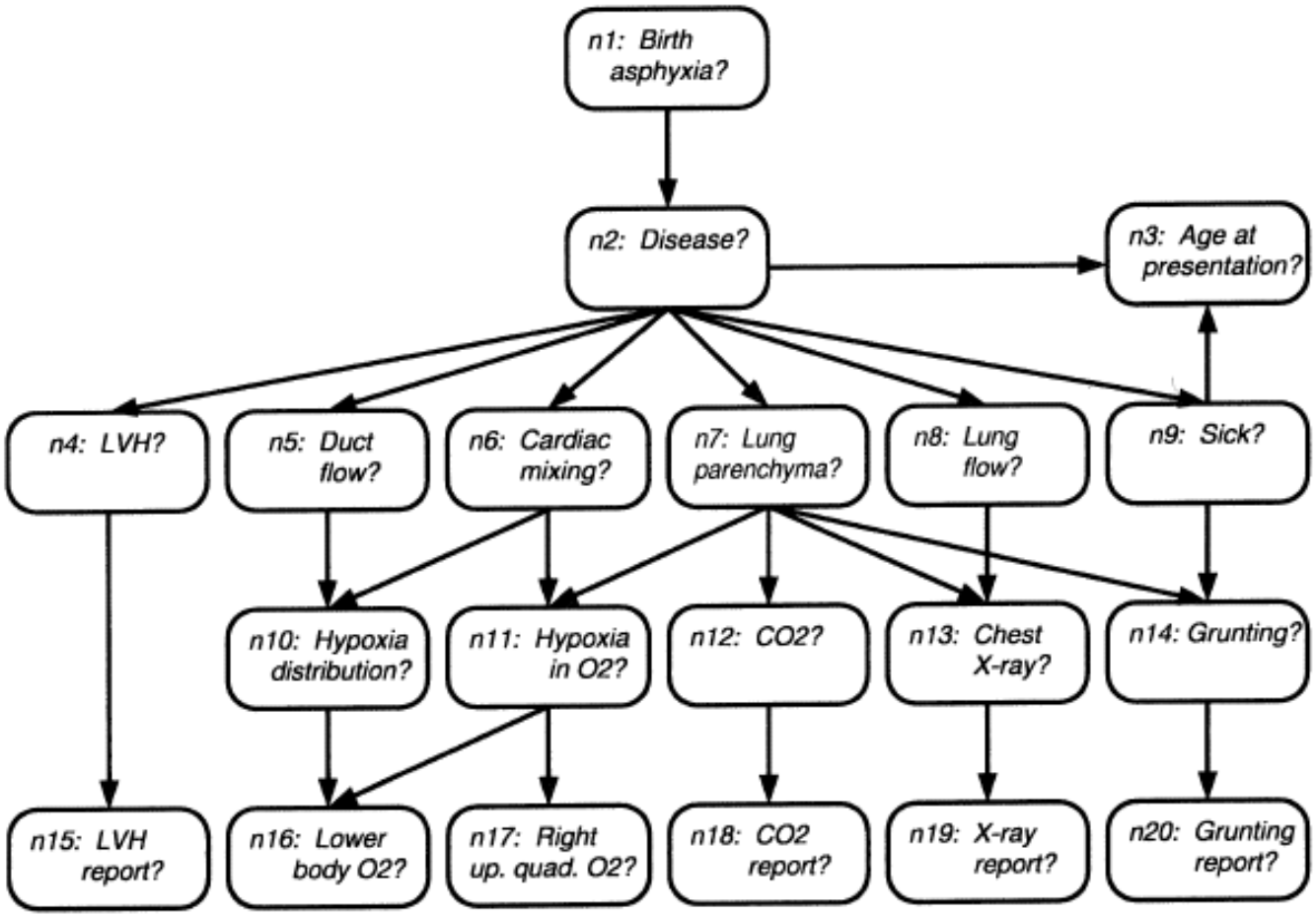


FIGURE 3.1. The CHILD network: Directed acyclic graph representing possible diseases that could lead to a blue baby.

Bayesian reasoning

Bayes' theorem:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

1. Model inference

$$p(\text{parameter}|\text{data}) = \frac{p(\text{data}|\text{parameter})p(\text{parameter})}{\sum_{\text{parameter}} p(\text{data}|\text{parameter})p(\text{parameter})}$$

with the *prior* $p(\text{parameter})$, the likelihood model $p(\text{data}|\text{parameter})$ and the posterior $p(\text{parameter}|\text{data})$.

2. Inference in the expert system

$$p(\text{disease}|\text{symptoms}) = \frac{p(\text{symptoms}|\text{disease})p(\text{disease})}{\sum_{\text{diseases}} p(\text{symptoms}|\text{disease})p(\text{disease})}$$

Notions from Graph Theory - I

- A *graph* $\mathcal{G}$ is a pair $\mathcal{G} = \{V, E\}$, where V is a finite set of *vertices*, also called *nodes*, of $\mathcal{G}$, and E is a subset of the set $V \times V$ of ordered pairs of vertices, called the *edges* or *links* of $\mathcal{G}$.
- If both ordered pairs (α, β) and (β, α) belong to E, we say that we have an *undirected* edge between α and β and write $\alpha \sim \beta$ (or $\alpha \sim_{\mathcal{G}} \beta$ to indicate the relevant graph $\mathcal{G}$). α and β are called to be neighbours. The set of *neighbours* of a vertex β is denoted by $ne(\beta)$.
- If $(\alpha, \beta) \in E$ but $(\beta, \alpha) \notin E$, we call the edge *directed*, and write $\alpha \to \beta$. We also say that α is a *parent* of β, $\alpha \in \mathrm{pa}(\beta)$, and that β is a *child* of α, $\beta \in \mathrm{ch}(\alpha)$.
- The *boundary* $\mathrm{bd}(\alpha)$ of a vertex α is the set of parents and neighbours of α, the boundary $\mathrm{bd}(A)$ of a subset $A \in V$ is the set of vertices in $V \setminus A$ that are parents or neighbours to vertices in A.
- The *closure* of A is given by $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$.

Notions from Graph Theory - II

- $\mathcal{G}_A = (A, E_A)$ is a *subgraph* of $\mathcal{G} = (V, E)$ if $A \subseteq V$ and $E_A \subseteq E \cap (A \times A)$. If $E_A = E \cap (A \times A)$, $\mathcal{G}_A$ is the subgraph of $\mathcal{G}$ induced by the vertex set A.
- A graph is called *complete* if every pair of vertices is joined. A complete subgraph is called a *clique*.
- A *path* of length n from α to β is called a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$.
- Let A, B, S disjoint subsets of V. Then S separates A from B if any path from A to B goes through S.
- An *n-cycle* is a path of length n with the modification that the endpoints are identical. We say that a graph is *acyclic* if it does not possess any cycles. A directed graph, which is acyclic is called a *directed acyclic graph (DAG)*.

Directed acyclic graphs - DAG

- Contains only directed links (arrows) and no cycles.
- Nodes are random variables, the links denote statistical dependencies. Note the difference to the representation of Markov processes where the nodes were states of a random variable.
- A probability distribution P admits a *recursive factorization* according to the graph $\mathcal{G}$ if

$$p(x_V) = \prod_{v \in V} p(x_v | x_{pa(v)}) \, .$$

- Interpretation as Bayesian network, belief networks, causal networks or generative model depends on the interpretation of the conditional probabilities $p(x_v|x_{pa(v)})$, the transition kernels.
- Causal networks (Pearl, 2000): The transition kernels are interpreted as mechanisms, which allow to study the effect of *interventions*. The kernel $p(x_v|x_{pa(v)}$ has to be stable under interventions which do not involve x_v.

Undirected graphs

- Undirected graphs contain only undirected links.
- If a undirected graph $\mathcal{G}$ is used as a graphical model the probability distribution factorizes according to

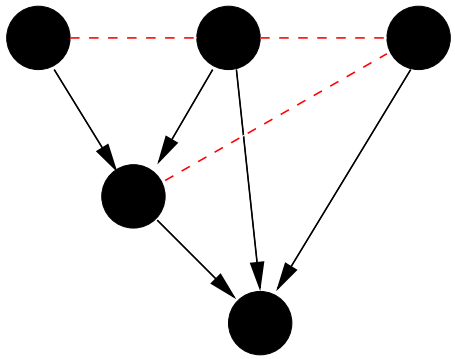
$$p(x_V) = \prod_{C \in \mathcal{C}} a_C(x_C)$$

  with $\mathcal{C}$ denoting the set of cliques (complete subgraphs) of $\mathcal{G}$.
- The $a_C(x_C)$ are also called potentials.
- This kind of graphical models is also known as Markov random fields.
- Examples: Gibbs distributions from statistical mechanics, e.g. Ising model with next-neighbour interaction

$$p(x) = \frac{1}{Z} \exp\left(-\beta\left(-\frac{1}{2}\sum_{i \sim j} x_i x_j\right)\right)$$

Form the directed to the undirected graph - the "moral" graph

- Let $\mathcal{G} = (V, E)$ be a DAG. The moral graph $\mathcal{G}^{(m)} = (V, E^{(m)})$ is defined as follows:
  1. "Marrying" parents, i.e. introducing additional undirected edges between any two nodes with a common child.
  2. Ignoring directions.



- If P admits a recursive factorization according to a DAG $\mathcal{G}$ then it also factorizes according to the undirected Graph $\mathcal{G}^{(m)}$, because the sets $\{v\} \cup \mathrm{pa}(v)$ are complete subsets, cliques, in $\mathcal{G}^{(m)}$.

Markov properties for undirected graphs

Markov properties of a probability distribution P with respect to a graph $\mathcal{G}$:

Pairwise (P): For any pair (α, β) of non-adjacent vertices:

$$\alpha \perp\!\!\!\perp \beta | V \setminus \{\alpha, \beta\}$$

Local (L): For any vertex $\alpha \in V$: $\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) | \mathrm{bd}(\alpha)$

Global (G): For any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $\mathcal{G}$

$$A \perp\!\!\!\perp B | S \, .$$

Factorisation (F): For $\mathcal{C}$ denoting the set of all cliques of $\mathcal{G}$

$$p(x_V) = \prod_{C \in \mathcal{C}} \psi_C(x_C); \, .$$

$(F) \Rightarrow (G) \Rightarrow (L) \Rightarrow (P)$, but $(P) \Rightarrow (F)$ only if all state space are discrete and the density P is strictly positive.

Markov properties and d-Separation in DAG's

- A *trail* in a DAG $\mathcal{D}$ is a path in the undirected version $\mathcal{D}$, i.e. where the directions of the arrows is ignored.
- A trail π from a to b in $\mathcal{D}$ is said to be *blocked* by S if it contains a vertex $\gamma \in \pi$ such that either
  - $\gamma \in S$ and arrows of π do not meet head-to-head at γ, or
  - γ and all its descendants are not in S, and arrows of π meet head-to-head at γ.
- Two subsets A and B are said to be **d-separated** by S if all trails from A to B are blocked by S.
- A and B are separated by S in $\mathcal{G}^{(m)}_{\mathrm{An}(A \cup B \cup S)}$ is equivalent to S d-separates A from B. Thus d-separation for the case of the directed graph is equivalent to the global markov property in the undirected case.

What are graphical models good for?

- Propagating evidence for instance in expert systems
- Mathematical theory of causality and the identification of causal effects (Pearl 2000).
- General framework for learning parameters and structure from data including hidden markov models (HMM), generalzed linear models (GLM) or neural networks
- Next lecture: Application to time series - Granger causality, Transfer entropy and other measures of interaction