
Lonnie Chrisman
September 18, 2002

1 Overview

This supplement contains some of the detailed technical details of the probabilistic model, algorithm, and experiments contained in the paper:


This document is not a stand-alone document – the original paper contains motivations, descriptions, and related references. The information in the supplement is provided for those who want to need the gory details.

2 Causal Structure

A causal model on a set of variables, $\mathbf{x} = \{x_1, \ldots, x_n\}$ is defined by $(M, \theta_M)$, where $M$ is a causal structure and $\theta_M$ is a collection of parameters for $M$. More precisely, $M$ is a set of directed links between pairs of variables, defining a directed graph that we require to be acyclic. The parameters are used to define a joint probability density over $\mathbf{x}$, $P(\mathbf{X}|M, \theta_M)$. We factor the joint density into a product of local densities, such that

$$P(\mathbf{X}|M, \theta_M) = \prod_{i=1}^{n} P(X_i|X_{\text{par}_M}(x_i), M, \theta_M)$$

(1)

Note that $X_{\text{par}_M}(x_i)$ denotes the joint assignment of values to all of the parents of $x_i$. The factorization in (1) is the standard factorization employed by
Bayesian networks, in which the full joint distribution is factored into local models specifying the density of each variable conditioned on its parents in $M$.

Within the space of possible Bayesian network representations, there are many ways to define a set of parameters, $\theta_M$, and to translate these into local models, $P(X_i|X_{\text{par}_M(x_i)}, M, \theta_M)$. A parameterization is a specific choice of parameters and a specific way to translate these to a local conditional density. Our PSB-2003 paper used one particular parameterization, which is spelled out in detail in what follows.

2.1 Variable Values

We assume that each variable can take on one of three discrete values: + (up-regulated), 0 (unchanged), and − (down-regulated). These values are interpreted to mean that the value for the variable changed (or was changed) in the indicated fashion relative to a reference condition. In other words, the values indicate a relationship between two situations, rather than an absolute value for the variable in a single situation. We will call a particular comparison between one particular situation and one particular reference situation a scenario. If we reverse the roles of the two situations of a scenario, we get a new scenario, which we term the reciprocal scenario. Whenever $X_i = +$ in the original scenario, we will have $X_i = −$ in the reciprocal scenario, and vice versa.

A model $(M, \theta_M)$ is considered to be a valid depiction of an underlying biological system over a certain scope of physiologic conditions, called the scope of applicability of the model. A scenario is considered to be within the scope of the model when both its situation and its reference situation are within the scope of the model. It therefore follows that a whenever a scenario is within the scope of a model, its reciprocal scenario is also within the scope of the model. Based on this, we introduce the Axiom of Symmetry, requiring that any predictions made by a model for a scenario must be consistent with the predictions the same model makes for the reciprocal scenario. The Axiom of Symmetry results in a substantial reduction in the free parameters.

2.2 Local Models

A basic problem with “non-parametric” local models is that the number of free parameters required to represent $P(X_i|X_{\text{par}_M(x_i)}, M, \theta_M)$ increases multiplicatively with the number of parents. When boolean variables are used, the noisy-OR is often employed to avoid this pitfall. The noisy-OR is not directly applicable in our situation, since we have 3-valued variables, but we can employ a generalization of the noisy-OR, called a mixture model. Let $\text{par}_M(x_i) = \{y_1, \ldots, y_m\}$,
then the local model is given by

\[
P(X_i = a | Y_1 = b_1, \ldots, Y_m = b_m, M, \theta_M) = w_0 \theta_{X_i = a} + \sum_{j = 1}^{m} w_j \theta_{X_i = a | Y_j = b_j}
\]

s.t.

\[
b_j \neq 0
\]

(2)

When a parent variable does not change, i.e., \( Y_j = 0 \), then we take it to have no influence on the change of the child, hence the extra condition that \( b_j \neq 0 \) in the sum - i.e., we only mix in those parents that change. Implicit also is that we only mix in those parents that are in \( M \). The same \( \theta_{X_i = a | Y_j = b_j} \) parameter applies to any model containing an edge from \( Y_j \) to \( X_i \), so that Equation (2) adapts to any \( M \) as the algorithm explores the space of possible model structures. The term in front of the sum adds in a basal or ambient source of uncertainty.

Although Equation (2) contains \( 7m + 4 \) parameters,\(^1\) the Axiom of Symmetry and the requirement that certain parameters sum to 1 reduces this number of parameters to \( 3m + 2 \) per node (or \( 3m + 2n \) all together, taken \( m \) to be the number of plausible links, and \( n \) the number of variables). In particular, we adopt these constraints:

\[
\sum_{j = 0}^{m} w_j = 1
\]

\[
\sum_{a \in \{+,-,0\}} \theta_{X_i = a | Y_j = b_j} = 1
\]

\[
\sum_{a \in \{+,-,0\}} \theta_{X_i = a} = 1
\]

\[
\theta_{X_i = a} = \theta_{X_i = -a}
\]

\[
\theta_{X_i = a | Y_j = b_j} = \theta_{X_i = -a | Y_j = -b_j}
\]

where \(-a\) is the reciprocal value (e.g., when \( a = + \), \(-a = -\), etc. The latter two constraints follow from the Axiom of Symmetry. These constraints mean that we need only keep track of \( \theta_{X_i = + | Y_j = +} \), \( \theta_{X_i = - | Y_j = b_j} \), and an unnormalized version of \( w_j \) for each potential parent \( Y_j \), and \( \theta_{X_i = +} \) and an unnormalized \( w_0 \) for each node. All other parameters follow from the constraints.

### 2.3 Meta-Assessments

We do not assume that the parameters \( (w_j, \theta_{X_i = a | Y_j = +}, \text{ and } \theta_{X_i = a}) \) are known or directly assessed by the end-user. Instead, the program uses Dirichlet con-

\(^1\)There are \( m + 1 \) mixture weight parameters \( w_j \), three ambient noise parameters, \( \theta_{X_i = a} \) for \( a \in \{+,-,0\} \), and \( 6m \) parent influence parameters, \( \theta_{X_i = a | Y_j = b} \) for \( a \in \{+,-,0\} \) and \( b \in \{+, -\} \)
stants, $a_{w_j}$, $a_{X_i|Y_j}$, and $a_{X_i}$. In most cases, the default Dirichlet constants supplied by the program are used, but in some cases knowledge can be represented by altering the Dirichlet parameters.

First, a quick review of the Dirichlet pdf is warranted before continuing. Let $\theta_1, \ldots, \theta_i$ be a set of parameters that sum to 1. A Dirichlet density over these parameters, specified by the positive constants $a_1, \ldots, a_i$, is given by

$$f(\theta_1, \ldots, \theta_i) = \frac{\Gamma \left( \sum_{i=1}^{l} a_i \right)}{\prod_{i=1}^{l} \Gamma (a_i) } \prod_{i=1}^{l} \theta_i^{a_i-1}$$

when $\sum_{i=1}^{l} \theta_i = 1$, $f(\cdot) = 0$ otherwise.

For each group of parameters in our model that must sum to 1, we impose a Dirichlet prior over the values of these parameters. Multiplying all these Dirichlet priors together, we have a joint distribution $P(\theta|\alpha)$, which is the second term in Equation (2) of the original paper. Note that this does not depend on $M$.

The treatment of our $w_j$ parameters deserve some attention since the $w_j$’s used change as our algorithm explores different model structures. In the case of $w_j$, we actually have two levels of normalization. First, we have a Dirichlet prior over the $w_j$’s corresponding to all plausible incoming links to a node. It is this Dirichlet density that we use for $P(\theta|\alpha)$. However, in Equation (2), we normalize again over only those incoming links that are in $M$. 

4