Computational experiment design (active learning)

Estimating complex models from data typically requires more data than we have. One way to minimize the amount of data needed either to estimate a model or distinguish between competing models is by optimizing the selection of experiments to carry out. We consider here automating the selection process. Automated experiment design or active learning can be used in many biological contexts, including probe design, which factor to profile, which gene to knockout, and so on. We will focus here for simplicity on gene deletions to discriminate among causal Bayesian network models.

Problems to solve

There are many problems that we have to solve in order to automate the selection of experiments. To begin with we have to narrow our focus to a set of hypotheses that we can entertain computationally. The difficulty here is that we have to be able to formally search this set and be able to eliminate hypotheses that are unlikely to explain the data generated so far. We also have to explicitly define the set of possible experiments. This set is typically relatively small because we need to be able to tell for each pair (hypothesis, experiment) what we would expect to see as the outcome if the hypothesis were correct and we carried out the experiment. The outcomes can be probabilistic but have to be well-defined. Lastly, we have to associate a value to changes in our beliefs over hypotheses so that potential experiments can be ranked by their expected value. Computational experiment design is indeed based on “expected value of information” calculations.

We will now outline the key computational problems we have to solve:

Problem 1: Representation of hypotheses. This problem is typically difficult for two reasons. First, the space of possible hypotheses we’d like to consider is very large. Second, the hypotheses have to provide quantitative predictions that can be evaluated based on the available experiments. We will consider here only Bayesian network structures (initially including parameters) as different hypotheses.

Problem 2: Definition of possible experiments. We will only consider gene deletions since these experiments, as interventions, are easy to use in the context of Bayesian networks (see earlier lectures).

Problem 3: Predicted outcomes. If we delete a gene we can expect its expression to be relatively low. But we have to also fill in likely values for the other variables as if we had carried out the experiment. The problem is that the predicted values for the other
variables in response to an experiment are likely to strongly depend on the hypothesis. We also don’t know which one of the hypotheses is correct. We therefore have to be able to generate values for the other variables based on each possible hypothesis and weight the samples according to how much we currently believe in each hypothesis. We will untangle this a bit in the context of the simple example.

**Problem 4:** Posterior over hypotheses. Once we have sampled a predicted outcome for a specific hypothetical experiment, we need to evaluate the posterior probability over the hypotheses that would have resulted had we seen the sampled values as observed data. The “value of information” associated with the experiment is now a function of this revised distribution over hypotheses. For example, we could define it as the number of bits gained about the identity of the correct hypothesis.

**Problem 5:** Determining the expected value of information for each possible experiment. Since the predicted outcomes are probabilistic, we have to repeatedly sample (or explicitly sum over the) possible outcomes to evaluate the “expected value of information” for each experiment.

**Example context and solutions to the problems**

**Problem 1:** Consider the following two graphs, G1 and G0, relating the expression of $f$, a gene corresponding to a potential transcription factor, and another gene $g$. Graph G1 postulates that $f$ (directly) influences the transcription of $g$ (e.g., as a DNA binding activator though the mechanism is not articulated in the model), while G0 assumes the expression levels of the two genes are controlled by independent means.

\[
\begin{align*}
G1 : & \quad f \rightarrow g \\
G0 : & \quad f \not\rightarrow g
\end{align*}
\]

We will begin by assuming that both $f$ and $g$ are binary variables, i.e., taking values in $\{0, 1\}$. Moreover, we will initially specify the hypotheses as graphs with specific parameters:

\[
\begin{align*}
G1 : & \quad P_1(f) : \begin{array}{c|c|c}
0 & 0.5 & 0.5 \\
1 & 0.5 & 0.5 \\
\end{array}, & P_1(g|f) : \begin{array}{c|c|c|c}
0 & 0 & 0.9 & 0.1 \\
1 & 0.1 & 0.1 & 0.9 \\
\end{array} \\
G0 : & \quad P_0(f) : \begin{array}{c|c|c}
0 & 0.5 & 0.5 \\
1 & 0.5 & 0.5 \\
\end{array}, & P_0(g) : \begin{array}{c|c|c}
0 & 0.5 & 0.5 \\
1 & 0.5 & 0.5 \\
\end{array}
\end{align*}
\]

where the subindexes refer the model so that, e.g., $P_0(f)$ always refers to the probabilities associated with graph G0. Note that the models are marginally consistent in the sense that
$P_1(f)$ from $G_1$ equals $P_0(f)$ from $G_0$. Similarly, $\sum_{f=0,1} P_1(f) P_1(g|f) = P_0(g)$. The models could therefore have been estimated from the same data.

We assume that we have no prior reason to prefer one over the other: $P(G_0) = P(G_1) = 1/2$.

**Problem 2:** There are only two possible deletion experiments we can carry out: $\text{set}(f = 0)$ or $\text{set}(g = 0)$. The notation is used to distinguish observed values such as $g = 0$ (hypothetical or real) from interventions $\text{set}(g = 0)$.

**Problem 3:** Let’s start with the experiment $\text{set}(g = 0)$. We can determine possible values for $f$ in response to this experiment by assuming that one of the models is correct (we just don’t know which one). By intervening on $g$ the resulting graphs for the two models are

- $G_1$:
  - $f$ (set $g = 0$)
  - $g$ (set $g = 0$)

- $G_0$:
  - $f$
  - $g$

Thus if $G_1$ is correct, then $f \sim P_1(f)$, and if $G_0$ is correct, then $f \sim P_0(f)$ in response to the experiment. Here $f \sim P_1(f)$ means that the value of $f$ is sampled using the probability table $P_1(f)$. Since we don’t know which one of the models is correct, $f$ has to be sampled from a mixture distribution

$$P(f|\text{set}(g = 0)) = P(G1)P_1(f) + P(G0)P_0(f)$$

In our case this is just the uniform distribution $P(f = 0|\text{set}(g = 0)) = 0.5$.

When we set$(f = 0)$, the two models are

- $G_1$:
  - $f$ (set $f = 0$)
  - $\text{set}(f = 0)$
  - $g$

- $G_0$:
  - $f$
  - $\text{set}(f = 0)$
  - $g$

and we would sample responses for $g$ as follows. If $G_1$ is correct, then $g \sim P_1(g|f = 0)$ and if $G_0$ is correct, then $g \sim P_0(g)$. The resulting mixture distribution over $g$ is

$$P(g|\text{set}(f = 0)) = P(G1)P_1(g|f = 0) + P(G0)P_0(g), \quad \frac{g = 0}{g = 1} = \begin{bmatrix} 0.7 & 0.3 \end{bmatrix}.$$

**Problem 4:** The posterior probabilities we have to evaluate are exactly those typical for mixture models. For example, if we $\text{set}(g = 0)$ then we have to determine $P(G_1|\{f, \text{set}(g = 0)\})$.
for all values of responses \( f \). More precisely,

\[
P(G_1|\{f, \text{set}(g = 0)\}) = \frac{P_1(f)P(G_1)}{P(f|\text{set}(g = 0))} = \frac{P_1(f)P(G_1)}{P(G_1)P_1(f) + P(G0)P_0(f)}
\]

For example,

\[
P(G_1|\{f = 0, \text{set}(g = 0)\}) = \frac{0.5 \cdot 0.5}{0.5 \cdot 0.5 + 0.5 \cdot 0.5} = 0.5
\]

Since there are only two possible hypotheses, \( P(G_0|\{f, \text{set}(g = 0)\}) = 1 - P(G_1|\{f, \text{set}(g = 0)\}) \).

Similarly, when we \( \text{set}(f = 0) \), we have to evaluate

\[
P(G_1|\{\text{set}(f = 0), g\}) = \frac{P_1(g|f = 0)P(G_1)}{P(g|\text{set}(f = 0))} = \frac{P_1(g|f = 0)P(G_1)}{P(G_1)P_1(g|f = 0) + P(G0)P_0(g)}
\]

where, for example,

\[
P(G_1|\{\text{set}(f = 0), g = 0\}) = \frac{0.9 \cdot 0.5}{0.9 \cdot 0.5 + 0.5 \cdot 0.5} = 0.64
\]

as the response \( g = 0 \) is more consistent with \( G_1 \) that represents \( f \) as an activator.

**Problem 5:** We will use *information gain* to assess the value of each experiment. Information gain in our context corresponds to the expected reduction of uncertainty about the hypotheses due to an experiment. Put another way, it is the difference between the uncertainty over hypotheses before and after a hypothetical experiment. The uncertainty after an experiment depends on the predicted responses that are probabilistic. We therefore can only talk about expected reduction, averaged over predicted responses.

We will use the Shannon entropy as a measure of uncertainty. You can think of entropy as the expected number of binary questions one would need ask in order to determine the value of a random variable. Let \( h = 0, 1 \) refer to the hypotheses, then

\[
Entropy(h) = - \sum_{h=0,1} P(G_h) \log_2 P(G_h) = 1 \text{ bit}
\]

which makes sense since our prior over the two hypotheses is uniform. Similarly, the entropy of hypotheses after \( \{\text{set}(f = 0), g = 0\} \) (we \( \text{set}(f = 0) \) and observe a response \( g = 0 \)) is

\[
Entropy(h|\{\text{set}(f = 0), g = 0\}) = - \sum_{h=0,1} P(G_h|\{\text{set}(f = 0), g = 0\}) \log_2 P(G_h|\{\text{set}(f = 0), g = 0\})
\]

\[
= -(1 - 0.64) \log_2 (1 - 64) - 0.64 \log_2 0.64
\]
The response $g$ is not known so we have to settle for the expected entropy after $set(f = 0)$, weighted by our mixture distribution of possible values for $g$

$$\sum_{g=0,1} P(g|set(f = 0)) \text{Entropy}(h|\{set(f = 0), g\})$$

The information gain is now the difference between the entropy before and the expected entropy after an experiment:

$$\text{Gain}(set(f = 0)) = \text{Entropy}(h) - \sum_{g=0,1} P(g|set(f = 0)) \text{Entropy}(h|\{set(f = 0), g\}) > 0$$

$$\text{Gain}(set(g = 0)) = \text{Entropy}(h) - \sum_{f=0,1} P(f|set(g = 0)) \text{Entropy}(h|\{f, set(g = 0)\}) = 0$$

So we would select $set(f = 0)$ as the experiment to carry out. In our simple context you could have immediately inferred that $set(g = 0)$ is useless for distinguishing G1 and G0 as the two models make identical predictions about $f$ in response to $set(g = 0)$. 