Causal Explorer FAQ (version from 8/17/2005)

Question 1: Which algorithm do you recommend to use to find a Markov Blanket, and how much sample do I need?

Answer: We recommend using *HITON_MB* to find Markov Blanket (MB) of a variable. We also recommend using *HITON_PC* that outputs a set of parents and children and occasionally provides a very good approximation of the Markov Blanket very quickly.

For discrete data you will need to specify a parameter k (maximum size of the conditioning set) that is large enough according to the following heuristic: "5 sample instances per independent cell in the multiway contingency table". For example, say our MB has 10 members/variables but it is feasible (because of the specific network connectivity) to establish that all non members are not in the MB by conditioning on up to 3 variables each time. Suppose for illustrative purposes that all non-target variables are ternary (i.e., have 3 values) and that the target variable is binary (i.e. has 2 value). This means that the algorithm will need at most $2*3^4$ = 162 cells times 5 instances = 810 instances. Let's examine some possibilities for calling the algorithm and what is the theoretically expected outcome:

HITON_MB called with k	# of instances	outcome
>2	<810	true MB + false positives
>2	>=810	true MB
<3	<810	true MB + false positives
<3	>=810	true MB + false positives

Note: false positives are created because the algorithm will be unable to determine only MB members by conditioning on <3 variables even with large sample.

The above means that when using the discrete *HITON**, use a large enough k (as determined by the sample and expected size of the MB, as per our example). Occasionally we will experiment with smaller k to see if we get a small enough MB estimate faster than the settings implied by the general rule above.

For continuous data and the z-test: the algorithms do not perform automatic checks on the data and thus it is important to provide a k that is small enough for the algorithm to terminate fast and k that is large enough for good quality. Typically we use the largest k that for our data terminates in acceptable time.

Question 2: Which algorithm do you recommend to use to discover a Bayesian Network, and how much sample do I need?

Answer: We recommend using *MMHC* algorithm to discover a Bayesian Network. Use the same approach about specifying k (maximum size of the conditioning set) as described for the Markov Blanket algorithms (see answer to the Question 1).

Question 3: Which algorithm do you recommend to use to find a Markov Blanket when I have thousands of samples?

Answer: The IAMB* family will give you the true MB faster than HITON_MB and MMMB.