Determining the Number of Non-Spurious Arcs in a Learned DAG Model

Jennifer Listgarten and David Heckerman
Microsoft Research
One Microsoft Way
Redmond, WA 98052

In many application areas where graphical models are used and where their structure is learned from data, the end goal is neither prediction nor density estimation. Rather, it is the uncovering of discrete relationships between entities. For example, in computational biology, one may be interested in discovering which proteins within a large set of proteins interact with one another. In these problems, relationships can be represented by arcs in a graphical model. Consequently, given a learned model, we are interested in knowing how many of the arcs are real or *non-spurious*.

In our approach to this problem, we estimate and control the *False Discovery Rate* (FDR) [1] of a set of arc hypotheses. The FDR is defined as the (expected) proportion of all hypotheses (*e.g.*, arc hypotheses) which we label as true, but which are actually false (*i.e.*, the number of false positives divided by the number of total hypotheses called true). In our evaluations, we concentrate on directed acyclic graphs (DAGs) for discrete variables with known variable orderings, as our problem of interest (concerning a particular problem related to HIV vaccine design) has these properties.

We use the term *arc hypothesis* to denote the event that an arc is present in the underlying distribution of the data. In a typical computation of FDR, we are given a set of hypotheses where each hypothesis, i, is assigned a score, s_i (traditionally, a test statistic, or the p-value resulting from such a test statistic). These scores are often assumed to be independent and identically distributed, although there has been much work to relax the assumption of independence [2]. The FDR is computed as a function of a threshold, t, on these scores, $FDR = FDR(t)$. For threshold t, all hypotheses with $s_i \geq t$ are said to be significant (assuming, without loss of generality, that the higher a score, the more we believe a hypothesis). The FDR at threshold t is then given by $FDR(t) = E\left[\frac{F(t)}{S(t)}\right]$, where $S(t)$ is the number of hypothesis deemed significant at threshold t and $F(t)$ is the number of those hypotheses which are false, and where expectation is taken with respect to the true joint distribution of the variables. When the number of hypotheses is large, as is usually the case, one can take the expectation of the numerator and denominator separately: $FDR(t) = E\left[\frac{F(t)}{S(t)}\right] \approxeq \frac{E[F(t)]}{E[S(t)]}$. Furthermore, it is often sufficient to use the observed $S(t)$ as an approximation for $E[S(t)]$. Thus the computation of $FDR(t)$ boils down to the computation of $E[F(t)]$. One approximation for this quantity which can be reasonable is $E[F(t)] \approxeq E_0[F(t)]$, where E_0 denotes expectation with respect to the null distribution (the distribution of scores obtained when no hypotheses are truly significant), and it is this approach that we take. Note that the FDR is closely related to *positive predictive value* (PPV), where $PPV(t) = 1 - \frac{F(t)}{S(t)}$. That is, FDR is 1 minus expected PPV.

Applying this approach to estimating the number of non-spurious arcs in a given (learned) DAG model, we take as input a particular structure search algorithm $\boldsymbol{a}$ (which may have hyperparameters such as κ that control the number of arcs learned) and generalize $S(\cdot)$ and $F(\cdot)$ to be functions of $\boldsymbol{a}$. In particular, $S(\boldsymbol{a})$ is the number of arcs found by $\boldsymbol{a}$ and $F(\boldsymbol{a})$ is the number of those arcs whose corresponding hypotheses are false. As in the standard FDR approach, we use the approximation $E(S(\boldsymbol{a})) \approxeq N(D, \boldsymbol{a})$, where $N(D, \boldsymbol{a})$ is the number of arcs found by applying $\boldsymbol{a}$ to the real data D. In addition, we estimate $E_0(F(\boldsymbol{a}))$ to be $N(D^q, \boldsymbol{a})$ averaged over multiple data sets D^q, $q = 1, \ldots, Q$, drawn from a null distribution. That is, $FDR(\boldsymbol{a}) = E\left[\frac{F(\boldsymbol{a})}{S(\boldsymbol{a})}\right] \approxeq \frac{E[F(\boldsymbol{a})]}{E[S(\boldsymbol{a})]} \approxeq \frac{(1+\sum_{q=1}^{Q} N(D^q, \boldsymbol{a}))/Q}{N(D, \boldsymbol{a})}$. The addition of one to the numerator smooths the estimate of $E_0[F(\boldsymbol{a})]$ so as to take into account the number of random permutations performed.

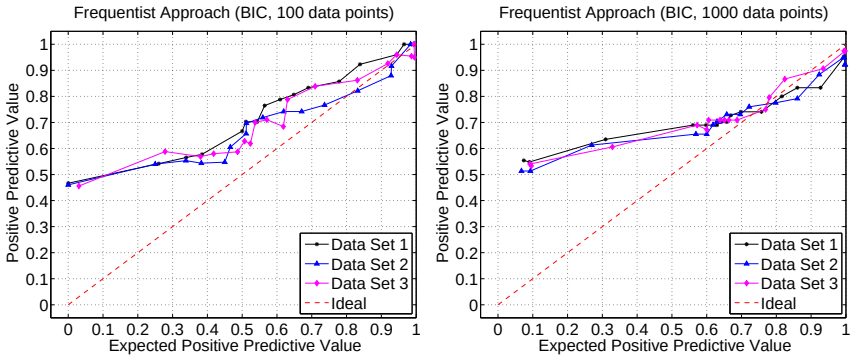

Figure 1: Evaluation of FDR estimates. Actual versus estimated expected PPV are shown. The dashed line denotes the idealized curve, where actual and estimated expected PPV are equal.

In our implementation of this approach, we assume that $\boldsymbol{a}$ has the property that it can be decomposed into independent searches for the parents of each node. Given this assumption, when we learn the parent set of a given node, we create the null distribution for that node by permuting the real data for the corresponding variable. This permutation guarantees that all arc hypotheses are false in the null distribution. The generation of these null distributions is computationally efficient as well as non-parametric, making them applicable to situations where the models learned are less representative of the data.

To determine whether our approximations are reasonable in practice, we draw samples from synthetic graphical models, run the algorithm above to compute the FDR, and then use the ground truth generating structure to measure the true FDR. For example, we evaluated the accuracy of our approach using data generated from the Alarm network [3], which contains 37 CPT-based nodes and 46 arcs.[1] From this model, we generated three data sets with sample size 100 and three with sample size 1000.

These models were learned by greedy structure search starting from the empty graph, where a single arc was added or deleted at each step of the search until the BIC score could not be increased. Depending on the setting of the structure prior, more or fewer arcs are learned during search, and thus we were able to generate a range of arcs learned (and hence a range of FDRs) for evaluation of our method.

The results are shown in Figure 1, which plots the expected positive predictive value against the actual positive predictive value according to the generating structure of the Alarm network. The expected PPVs (positive predictive values) plotted are simply $1 - FDR$. The curves tend to stay reasonably accurate for high PPVs, and then gradually peel away from the idealized curve, in a conservative manner. In real applications, the PPV range of interest is typically in the high end because one does not want an abundance of false hypotheses to pursue.

At the workshop, we will describe experimental results on other synthetic data sets, and on a real problem in HIV vaccine design.

Topic: graphical models Preference: oral/poster

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

[2] John D. Storey. A direct approach to false discovery rates. *J.R. Statist. Soc. B*, 64:479–498, 2002.

[3] I.A. Beinlich, H.J. Suermondt, R.M. Chavez, and G.F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, London, pages 247–256. Springer Verlag, Berlin, August 1989.

[1] We arbitrarily resolved the four non-compelled edges in the model by placing LVFailure before History, Anaphylaxis before TPR, PulmEmbolus before PAP, and MinVolSet before VentMach.