Discussion of “Learning Equivalence Classes of Acyclic Models with Latent and Selection Variables from Multiple Datasets with Overlapping Variables”

Jiji Zhang
Department of Philosophy
Lingnan University
Tuen Mun, NT, Hong Kong
jijizhang@ln.edu.hk

Ricardo Silva
Department of Statistical Science
University College London
Gower Street, London UK WC1E 6BT
ricardo@stats.ucl.ac.uk

In automated causal discovery, the constraint-based approach seeks to learn an (equivalence) class of causal structures (with possibly latent variables and/or selection variables) that are compatible (according to some assumptions, usually the causal Markov and faithfulness assumptions) with the conditional dependence and independence relations found in data. In the paper under discussion, Tillman and Spirtes (T&S) develop a constraint-based algorithm for learning causal structures from multiple, overlapping datasets. The basic setup of the problem is this: the variables of interest are not all measured at once in a single study. Instead there are several studies, each measuring a subset, which produce multiple datasets with overlapping variables. Assuming there is a common structure over the variables of interest (with possibly latent confounding variables and selection variables) that generated all the datasets, T&S's algorithm is designed to discover features of that structure by learning the features shared by all the causal structures that are compatible with all the datasets .

Unlike standard constraint-based methods, which assume an oracle of conditional independence that can respond to every query of whether two observed variables are conditionally independent given a set of observed variables, the algorithm described in T&S's paper allows the oracle to be incomplete: some variables may be jointly measured in none of the available datasets; as a result, the query of conditional independence concerning these variables cannot be answered. Thus the algorithm provides a solution to the problem of learning from (a broad class of) incomplete oracles. To be sure, the algorithm does not allow the oracle to be arbitrarily incomplete: with respect to the variable set of each individual dataset, the oracle is still complete. But it seems to be general enough to handle the problem of incomplete oracle in some other contexts. For example, one may worry about the power of conditional independence tests used to implement the oracle or about the computational complexity, and decide to only consult the oracle when the conditioning set contains no more than n variables (Spirtes, 2001). T&S's method seems also applicable in this case, which can be viewed as a situation in which there are multiple datasets, each containing a subset of the whole variable set with size $n + 2$. (In this case, though, we can simply use the aggregate dataset to test conditional independence, and do not need Fisher's method of meta-analysis.)

T&S assume that every dataset is generated from a common qualitative structure (though different variables get recorded in different datasets), but not necessarily with the same parameter values. In general, therefore, we have one underlying structure but multiple joint distributions generated from the structure. Moreover, each joint distribution is assumed to be faithful to the common structure, so that every conditional independence relation that holds in the joint distribution is entailed by the structure, regardless of the specific parameter values.

The flexibility of allowing different distributions for different datasets is obviously commendable. For one thing, it is simply more realistic to expect different datasets to follow different distributions, if, as is usually the case, they are collected in different studies or experiments with different designs. Moreover, the multiple-distribution setting may have an advantage with regard to the faithfulness assumption. As Pearl once remarked: “[It] has been suggested that causal discovery methods based solely on associations will find their greatest potential in longitudinal studies conducted under slightly varying conditions, where accidental independencies are destroyed and only structural independencies are preserved.” (Pearl, 2009, p.63) The idea is that even if one distribution happens to be unfaithful and implies an accidental conditional independence, the independence will in all likelihood disappear when different distributions are checked. In the present context, if multiple datasets contribute to testing whether X and Y are conditionally independent given Z, there is a good reason to ex-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

pect that even if the conditional independence happens to hold unfaithfully in one dataset, other datasets, if they are faithful with respect to this particular conditional independence query, will guard against the misleading statistical decision that X and Y are conditionally independent given Z. In their paper, T&S propose to use Fisher's method of meta-analysis to integrate different datasets in testing conditional independence. Our speculation is that this method is to some extent robust against occasional violations of faithfulness in multiple-distribution settings. We are curious to see some empirical results on this matter.

Let's now turn to the common-structure assumption. The idea is that there is a common DAG structure representing the data-generating process for each dataset. T&S allow the possibility of selection bias, so this underlying DAG may contain selection variables. A selection variable is a variable that remains constant in sampling; as a result, whatever dependence or independence found in the sample is conditional upon the variable (taking a certain value). In the simplest case, a selection variable matters for causal inference when two variables of interest both causally influence the selection variable, in which case conditioning on the selection variable (a collider) induces an association between the variables of interest that is not attributable to either direct causal influence between the variables or a common cause of the variables. What T&S need to assume is that all the datasets, despite coming from different studies or experiments, nonetheless share the exact same selection variables (perhaps even with the same values, though this is not necessary if linear Gaussian parameterization is assumed). This aspect of the assumption seems to us not very plausible. Considering that researchers, at least in normal cases, do not intentionally introduce selection bias in experimental design, it would take quite a coincidence to have the exact same selection bias in multiple experiments. Possible relaxation of the assumption of common selection bias is worth exploring.

More generally, there may be situations in which the common-structure assumption is known to be false. It might be known, for example, that one dataset comes from a controlled experiment, while others come from observational studies. Different experiments may also carry out interventions on different variables. How to learn from such datasets, whose underlying structures are overlapping but not identical, seems to be a natural follow-up question.

The contribution by T&S also provides a framework that can go beyond discovering a collection of Markov equivalence classes. Consider the problem of inferring the presence of latent variables, their relation to the observables, and the relation among themselves (Silva et al., 2006). For instance, suppose data for a natural process is generated according to the following causal structure

$$L \to X_1, L \to X_2, L \to X_3 \dots, L \to X_K$$

but where L is never observed, or even known to exist a priori. Under a variety of assumptions, such as faithfulness and linearity, several causal claims can be inferred from the marginal distribution of $\{X_1, \dots, X_K\}$ without further assumptions about the number of latent variables and their marginal distribution. Given the existence of many studies where observed variables are recorded with the intent of measuring latent concepts of interest (up to some measurement error) (Bartholomew et al., 2008), it is important to provide tools to unveil such relationships.

For a variety of reasons, such studies have overlapping variables. In medicine, psychology or social sciences this is particularly evident: each X_i might record an answer to a question probing a target latent trait of interest (say, propensity to anxiety disorder). Such a trait might be measured in a study that uses some but not all of the questions from a different study, while including a few more. Cudeck (2000) approaches this problem under the assumption of a given structure and common distributions. In principle, this could be done with standard missing data models, but Cudeck's main goal was to explore the possibility of combining assumptions and data to identify joint distributions of variables which were never observed together. The framework by T&S could be adapted to such a scenario, where different marginal distributions, augmented with latent variables, contain different independence constraints. Tests for independence should now be done in an indirect way by exploring assumptions such as linearity (Silva et al., 2006) — which again means we have a special type of limited information about the Markovian structure of the joint distribution. T&S already equip search algorithms with the possibility of working under incomplete independence information. This seems a promising starting point for such problems.

Concerning other families of constraints that could be exploited, T&S mention functional constraints with additive errors such as the one introduced by Hoyer et al. (2009). Such a framework provides a different approach for identifying causal structures, one that often results in much simpler equivalence classes. Since additive error models are not closed under marginalization, however, other issues will arise: what would that mean when one model has additive error, if the error might actually come from a variable observed in another dataset, where the relationship was non-additive?

Some closing comments on model-based approaches: T&S correctly point out that score-based methods using standard methods such as Structural EM (Friedman, 1998) might underperform in their scenario. In theory, however, this is true only to some extent: if one does not extend such approaches to deal with different distributions. There is no conceptual barrier against dropping this assumption, and the Bayesian framework is particularly suitable if one is willing to tie the parameters from the different submodels via a common prior. It is true that the computational cost should be much higher than the procedure implemented by T&S — but if datasets are individually of modest dimensionality, it might be doable in practice.

Moreover, it might not be necessary to fit a single model including all variables. One could exploit the link to penalized composite likelihood approaches (Varin and Vidoni, 2005) and design a score function that is a sum of different score functions over different subsets of variables. Even if the data is not identically distributed, again we see no fundamental barrier to extend the approach to this general setup — although coming up with a reasonable penalization function might not be trivial. From the point of view of search algorithms for optimizing structure, much of the machinery of combinatorial optimization could be exploited to provide a counterpart to (at least) T&S's Algorithm 2: one could optimize the penalized composite likelihood score by enforcing constraints such that the independence models over different subsets of variables agree on the overlapping sets. At least in principle such constraints could be added to standard constrained optimization solvers, and since we have a sound and complete calculus to generate such constraints (i.e., m-separation), the problem is well-defined. Recent work (Cussens, 2008, Jaakkola et al., 2010) has exploited this formulation in the context of learning single DAG structures. The problem is much harder here since scoring MAGs cannot be easily done by decomposing the likelihood function in a simple way. However, the different subsets of variables in which our problem is naturally split might end up being small enough in some important practical cases and a missing data formulation is not necessary.

It is a great pleasure to participate in a discussion paper at AISTATS, arguably the first machine learning conference to provide such an opportunity. Following a tradition that one of us has observed frequently at the meetings of the Royal Statistical Society, we would like to propose a vote of thanks on this paper and once again congratulate the authors for their contribution.

References

Bartholomew, D., Steele, F., Moustaki, I., and Galbraith, J. (2008). *Analysis of Multivariate Social Science Data, 2nd edition*. Chapman & Hall.

Cudeck, R. (2000). An estimate of the covariance between variables which are not jointly observed. *Psychometrika*, 65(4):539–546.

Cussens, J. (2008). Bayesian network learning by compiling to weighted MAX-SAT. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*, pages 105–112.

Friedman, N. (1998). The Bayesian structural EM algorithm. *Proceedings of 14th Conference on Uncertainty in Artificial Intelligence (UAI '98)*.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. *Neural Information Processing Systems 21 (NIPS 2008)*, pages 689–696.

Jaakkola, T., Sontag, D., Globerson, A., and Meila, M. (2010). Learning Bayesian network structure using LP relaxations. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, pages 366–373.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference, 2nd edition*. Cambridge University Press.

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. (2006). Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246.

Spirtes, P. (2001). An anytime algorithm for causal inference. *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics (AISTATS 2001)*, pages 213–221.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.