Research Statement

Ricardo Silva
Center for Automated Learning and Discovery
rbas@cs.cmu.edu

December 15, 2004

Interests

I am interested in automated knowledge discovery, in particular statistical and causal models with hidden, or unmeasured variables. I consider graphical models with hidden variables a key framework for machine learning. Such models are flexible enough to encompass large classes of probability distributions through mixture models, provide valuable tools for decomposing complex problems into simpler ones via hierarchical models, and are essential in causal analysis in representing hidden common causes, i.e. confounders.

Recent Work

My current research has been in automatically discovering the existence of hidden variables and in automatically discovering how to measure them accurately, especially when the goal is to examine the relationships among hidden variables.

During the past few years as a PhD student at Carnegie Mellon University, I have studied how to identify features of a latent variable model under weak structural assumptions. That is, how can one provide consistent criteria to decide which hidden variables exist? How can one decide if an observed variable measures a hidden variable, if one cannot directly evaluate conditional independencies that involve unmeasured variables?

One can accomplish this task by searching among models that predict certain features of the purely observable marginal distribution. Consider the following assumption: assume that certain constraints on the marginal are not a result of an accidental choice of parameters, but a consequence of the unknown graphical structure that has to be learned from data. For instance, assume that observed conditional independencies are not due to specific conditional probability tables, but hold if and only if they are entailed in the unknown graphical structure by d-separation (this is sometimes called the *faithfulness*, or *stability* assumption [SGS00, Pea00]).

We adopted similar assumptions with respect to rank constraints on the observed covariance matrix. Some combinations of rank constraints that are judged to hold in the population according to the data can only be generated by specific structures, and by searching among different structures that predict these constraints we can learn several features of the unknown graphical model. Results can be obtained even when non-linear relations among latents are allowed [SSGS03]. We are able to provide stronger guarantees in the fully linear case [SSGS04].

There are two questions that might arise at this point: why is this important? And how reliable is this procedure?

The importance of these results are two-fold: first, in many situations the structure of graphical models can be given a causal interpretation, and it is important to provide some guarantees about a structure generated by any learning procedure. It is not enough to fit the data: it is also important to report all other structures that fit it. The set of features common to all models that predict the observed constraints will form our causal model, and with the right structure and extra assumptions it is possible to predict the effect of interventions [SGS00, Pea00]. Sometimes directionality of causality can be learned from data. Sometimes the directionality is given by background knowledge, and latent variable models are used to learn regression functions when measurement error is present [CRS95]. Our method can also be seen as a way of finding instrumental variables to learn regression functions under measurement error. In simulation studies [SSGS04], our method worked particularly well, while factor analysis was of very limited value even considering that the simulations respected all the assumptions adopted by factor analysis. Even though causal models are difficult to evaluate when experimental data is not available, learning them from observational data is still of great importance: for instance, learning regulatory networks of gene expression is mostly meaningless without causal assumptions or massive experiments.

Another important application of our results is in probabilistic modeling: although several latent variable models and algorithms are generic enough to provide reasonable density estimation, those are usually either based on ad-hoc search procedures [ELFK00], or are constrained to have very few degrees of structural freedom such as no more than the number of latents and number of mixture components [GB99]. A more principle set of structural identification conditions can help the design of search procedures to achieve better density estimation. In [Sil04] we provide our first empirical results for this task. Much remains to be done, such as using our theoretical results as a guideline for more complex Bayesian search methods.

My interest in graphical models and latent variables has also influenced my work during a Summer internship at Clairvoyance corporation, Pittsburgh. My project was a graphical representation of *models of work*, such as modeling the flow of activities in a large organization with the goal of designing policies to increase productivity. Based on the literature of workflow mining [vdAW04], we developed a coherent probabilistic model and an interesting class of workflow graphs that obey structural constraints that are believed to hold in such processes, such as the *nesting* of activities. Latent variables are indirectly used to model measurement error. The first results are given in [SZS04].

Suggested future work

Currently I am looking at latent variable models with discrete observed variables and continuous latents, and how my framework can be used under this scenario. I am also developing a different search algorithm based on Bayesian scoring and on the identification conditions for latent variables developed in my previous work.

For the immediate future, I foresee some very interesting possibilities. Since I have always been interested in learning large knowledge bases, the idea of statistical matching [Ras02], a framework which goal is to make inferences about a joint distribution without any direct observations of the joint, is particularly attractive. There are already some results in graphical models concerning data fusion. For instance, David Danks at CMU has some theoretical results on building single Bayesian networks from multiple data sources [Dan02]. [Cud00] describes approaches for estimating the covariance of two variables that are not jointly observed by fitting latent variables models (factor analysis, in this case). Notice this is not the same as learning from relational data, since different data sources are not necessarily measured over the same individuals.

I will also continue to pursue the topic of consistent identification of latent structure. The work of [TP02] and [GM99] might provide new ideas in this area. In particular, a natural extension of my work is in how to adapt it to deal with time-series data. Under the assumption that the latent model has a fixed structure over a finite window of time as in a dynamic Bayesian network, several of the constraints that we have been using so far can be carried to the dynamic case.

I am also very interested in nonparametric Bayesian methods applied to latent variable models. Carrol et al. [CRC+04] present new identification results in Bayesian nonparametric regression with measurement error and instrumental variables, a problem whose goals that are closely related to my previous work. How much of such techniques could be adapted to the problem of learning the structure of latent variable models is a relevant topic.

Bayesian analysis can also contribute to causal analysis in case if one is willing to put priors on how weak an empirical dependency should be (e.g., a sample correlation coefficient) in order to imply a structural independence (e.g., no causal connection). The "strong faithfulness" family of assumptions of [ZS03] is a formal non-Bayesian way of incorporating these priors.

Finally, I have a variety of interests in other areas of graphical models. Although my interests in latent variable models concern primarily causality discovery and density estimation, dimensionality reduction and approximate inference are also topics I consider of great relevance. I am particularly curious about a new family of multinomial models currently in development by Thomas Richardson and Mathias Drton (personal communication) based on the independence models of [RS02], which generalizes directed and undirected graphs. I would like to explore the possibility of developing approximate inference algorithms for these types of graphical models.

References

[CRC+04] R. Carroll, D. Ruppert, C. Crainiceanu, T. Tosteson, and M. Karagas. Nonlinear and nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, pages 736–750, 2004.

[CRS95] R. Carroll, D. Ruppert, and L. Stefanski. *Measurement Error in Nonlinear Models.* Chapman & Hall, 1995.

[Cud00] R. Cudeck. An estimate of the covariance between variables which are not jointly observed. *Psychometrika*, 65:539–546, 2000.

[Dan02] D. Danks. Learning the causal structure of overlapping variable sets. *Discover Science: Proceedings of the 5th International Conference*, pages 178–191, 2002.

[ELFK00] G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering hidden variables: a structure-based approach. *Neural Information Processing Systems*, 13:479–485, 2000.

[GB99] Z. Ghahramani and M. Beal. Variational inference for Bayesian mixtures of factor analyzers. *NIPS*, 1999.

[GM99] D. Geiger and C. Meek. Quantifier elimination for statistical problems. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 1999.

[Pea00] J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.

[Ras02] S. Rassler. *Statistical Matching.* Springer, 2002.

[RS02] T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30:962–1030, 2002.

[SGS00] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Cambridge University Press, 2000.

[Sil04] R. Silva. New d-separation identification results for continuous latent variable models. *To be submmited*, 2004.

[SSGS03] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning measurement models for unobserved variables. *Proceedings of 19th Conference on Uncertainty in Artificial Intelligence*, pages 543–550, 2003.

[SSGS04] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *To be submmited*, 2004.

[SZS04] R. Silva, Jiji Zhang, and J. G. Shanahan. Probabilistic workflow mining. *To be submmited*, 2004.

[TP02] J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 2002.

[vdAW04] W. van der Aalst and A. Wejters. Process mining: a research agenda. *Computers and Industry*, 53:231–244, 2004.

[ZS03] J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, 2003.