# Research Statement

*Ricardo Silva*
Gatsby Computational Neuroscience Unit
rbas@gatsby.ucl.ac.uk

November 11, 2006

## 1 Philosophy

My work lies on the intersection of computer science and statistics. The questions I want to answer are of the following nature: how can machines learn from experience? This raises questions about statistical modeling, since the nature of a phenomenon is only observable through a limited set of measurements: the data. Rather than explicitly programming a computer to perform a particular task, *machine learning* uses data and statistical models to achieve intelligent behavior. The outcome can be observed in tasks as diverse as: predicting user preferences (movie ratings are fashionable these days[1]); filtering spam; adapting models of computer vision and speech recognition to new environments; improving retrieval of important documents; improving machine translation; and many others.

We can also turn the question around and ask instead how machines can be used in new methods of data analysis, and improve scientific progress. Standard statistical practice focuses on studies with a small number of variables and data points, but the increase in the amount of data that has been collected is evident. The need for analysing high dimensional measurements, and combining different sources of data, is pressing. Now the issue turns to finding proper computational approaches for building models from data, and providing novel techniques for exploration and analysis within more thorough studies.

In particular, my research addresses fundamental questions on learning with graphical models. More precisely, models with hidden (latent) variables. Such models are appropriate when the observed associations in our data are due to hidden common causes of our measured variables. This happens, for instance, if the observations are sensor data measuring atmospherical phenomena, medical instruments measuring biological processes, econometrical indicators measuring economical processes and so on. Reyment and Joreskog (1996) and Bollen (1989) provide an extensive list of examples of this class.

Graphical models are a powerful language for expressing conditional independence constraints, a necessity if one aims to model large dimensional domains (Jordan, 1998). Graphical models also provide a language for causal modeling, as required if one needs to compute the effects of interventions. Examples of interventions are medical treatments, genetic engineering, public policy issues such as tax cuts, and marketing strategies, among others (Spirtes et al., 2000; Pearl, 2000).

I believe that the best approach in solving a real problem lies in a careful statistical formulation of the question, identifying how to best use parametric and nonparametric statistical principles, which dependencies are necessary, which hidden variables could or should be used to model the observable phenomena, and finally, which computational methods should be applied. Although a crucial component of any machine learning solution, I do not believe algorithms should be the starting point of any learning framework: my philosophy is to write down which family of models should be the most appropriate for that domain, and only then concentrate on how to compute the desired predictions or model selection criteria. When computational limits are reached, one should approximate what is, to the best of our knowledge, the correct model.

[1] http://www.netflixprize.com

## 2 Recent research

I have applied my philosophy to solve a range of original real-world problems:

**Discovery of causal and probabilistic latent structure**

The problem of learning causal graphs from data has an additional challenge that does not exist in non-causal problems: namely, it is important to report not only a structure that explains the data, but all compatible structures. This identification problem only gets more difficult when unknown hidden variables are common causes of several of our observed variables: observable conditional independencies disappear, which might severely limit the usefulness of standard approaches for learning causal graphs.

Although there are principled procedures for learning causal structure that are robust to the presence of hidden variables, they are mostly concerned with the case where many conditional independencies still exist in our observable marginal distribution (Spirtes et al., 2000; Pearl, 2000). In several problems, however, the data consists of a large number of measurements that are indicators of an underlying latent process. Such indicators are strongly marginally dependent. Examples of such processes can be found in psychological studies: individuals need to answer several questions measuring a few latent psychological traits. Different questions might be measuring different aspects of the same latent variable, and therefore no conditional independencies exist among the measured items. Another class of examples can be found in the natural sciences, where raw data coming from a set of instruments provides different aspects of the same latent natural phenomena. For instance, atmospherical phenomena measured by sensors at different frequencies.

Traditionally, factor analysis and its variants have been used to model such problems (Reyment and Joreskog, 1996). However, such heuristic approaches are often based on artificial "simplicity" criteria to generate a particular solution. I have instead developed a formal theoretical approach for identifying non-trivial equivalence classes of causal latent variable models, and developed algorithms according to that theory. It is interesting to notice that, until recently, even textbooks considered (Bartholomew and Knott, 1999, p. 190) such a solution to be unattainable. Two conference papers contain the basic results of this project:

- Silva, R.; Scheines, R.; Glymour, C. and Spirtes P. (2003). "Learning measurement models for unobserved variables." 19th Conference on Uncertainty in Artificial Intelligence, UAI '03.
- Silva, R. and Scheines, R. (2005). "New d-separation identification results for learning continuous latent variable models." International Conference in Machine Learning, ICML '05.

A recent journal paper contains a more thorough review of the problem, extended results and empirical evaluation:

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

I have also adapted the principles used in causal discovery to the problem of density estimation using graphical models with latent variables. Although under this setup there is no more need to consider equivalence classes of models, the same identifiability results can be used to design a better search space for greedy algorithms in Bayesian learning:

- Silva, R. and Scheines, R. (2006). Bayesian learning of measurement and structural models. 23rd International Conference on Machine Learning, ICML '06

A different take on the problem, within the data mining perpective, allows us to approximate the solution in large dimensional discrete data by focusing on generating submodels. Those submodels can be interpreted as causal association rules:

- Silva, R. and Scheines, R. (2006). "Towards association rules with hidden variables." 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD '06.

**Workflow analysis of sequential decision making**

Most large social organizations are complex systems. Every day they perform various types of processes, such as assembling a car, designing and implementing software, organizing a conference, and so on. A process is a set of tasks to be accomplished, where every task might have pre-requisites within the process that have to be fulfilled before execution. Such a process follows a sequence of decisions on which a step has to be evaluated and the next action, chosen.

Modeling the sequential execution of plans has been object of study in probabilistic and causal modeling for years. For instance, identification conditions for estimating causal effects in longitudinal studies have been given graphical model treatments (Pearl, 2000, Chapter 4).

Our problem here shares a few similarities with this setup, but it has its own particular assumptions and goals. We are interested in discovering the structure of sequential decision making that happens in such organizations using data that records the time of such activities. For instance, data that is recorded both in terms of structured databases in a workflow system, or in unstructured text data (e.g., e-mail communication).

Even if an organization already follows its own normative workflow model, learning such a model from data is also a way of verifying if said norms are being respected. Other applications include monitoring processes (is a set of activities being executed the way it should be?), outlier detection (is this particular instance being executed in a unlikely way?) and policy making (assuming causal semantics for a workflow, what will happen if I intervene in a particular stage of a process?)

Unlike traditional time-series processes, workflow processes correspond to several chains of actions that might happen in parallel. For instance, the process of manufacturing a car includes subprocesses of manufacturing individual parts, which can happen independently. However, there is a point in the process where such individual subtasks have to be synchronized (e.g., when parts have to be added in a particular order to the chassis). Learning points of parallelization and synchonization is part of the inference problem. Latent variables also play a role when we consider models for measurement error, i.e., when tasks are not properly recorded in the respective database.

Our current results in workflow modeling can be found in the following reference:

- Silva, R.; Zhang, J. and Shanahan, J. G. (2005). "Probabilistic workflow mining." Knowledge Discovery and Data Mining, KDD '05.

This work also resulted in a recently approved patent submitted to the U.S. Bureau of Patents.

**Bayesian inference for mixed graph models**

From a historical perspective, the contributions of Sewall Wright can be considered the starting point on the modern development of graphical models (Wright, 1921). From the beginning there was a need to distinguish between symmetric and asymmetric dependencies that connect given random variables. A common motivation is the need to distinguish between the asymmetric notion that $A$ causes $B$ from the notion that $A$ and $B$ have a hidden common cause. Mixed graphs (Richardson, 2003; Richardson and Spirtes, 2002) are a family for representing such a mixed type of dependencies within a single graph. Other symmetric/asymmetric graphical languages such as chain graphs are not in general suitable for this task (Richardson, 1998). Even if a model has explicit hidden variables, such as those generated by the discovery procedures of my previous contributions, one still has to decide if the relation between two dependent hidden variables is symmetric or asymmetric. Latent variable models do not avoid the necessity for mixed graph representations.

A mixed graph only encodes qualitatively conditional independencies. To specify a probabilistic model, it is necessary to specify a parameterization for a distribution that is Markov with respect to the graph. Gaussian mixed graph models are very common in several fields such as social sciences and econometrics (Bollen, 1989), where they are known as "structural equation models." To a lesser extent, they can also be found in biological domains (Shipley, 2002).

For Bayesian inference, one has also to specify priors for such parameters and a way of computing posteriors. Proper Bayesian treatment of Gaussian mixed graph models has remained elusive. Some solutions require the specification of improper priors and appeal to rejection sampling algorithms (Scheines et al., 1999).

Other solutions artificially insert extra hidden variables as surrogates for hidden common causes (Dunson et al., 2005), but this adds (possibly severe) bias (Richardson and Spirtes, 2002).

I have developed a sound, efficient way of performing Bayesian inference for Gaussian mixed graph models. This allows for proper priors, adds no extra bias, and does not require rejection or importance sampling. This procedure is described in:

- Silva, R. and Ghahramani, Z. (2006). "Bayesian inference for Gaussian mixed graph models." 22nd Conference on Uncertainty on Artificial Intelligence, UAI '06

Just recently, we have finished some extensions that partially provide a formulation for discrete and non-parametric models:

- Silva, R. and Ghahramani, Z. (2006). "Bayesian inference for discrete mixed graph models: normit networks, observable independencies and infinite mixtures." `http://www.gatsby.ucl.ac.uk/~rbas`

**Novel issues on exploratory data analysis for relational data**

Building predictive models is traditionally the main focus of machine learning (ML). However, there are many opportunities for ML methods in exploratory data analysis. For instance, many approaches for causal discovery, in practice, should be seen as exploratory data analysis methods: they provide evidence that is entailed by data and prior knowledge concerning possible causal pathways, indicate which extra information is needed in order to distinguish between equivalent models, and which experiments are more promising in order to unveil the desired information.

Besides processing causal information, I have been interested on evaluating similarities between relational structures. In particular, I have developed a new view of probabilistic analogical reasoning for identifying interesting subpopulations in a relational domain.

- Silva, R.; Heller, K. and Ghahramani, Z. (2006). "Analogical reasoning with relational Bayesian sets." `http://www.gatsby.ucl.ac.uk/~rbas`

The gist of the idea is as follows: imagine a set of relations. For simplicity, imagine pairs of linked objects. These could be pairs of papers $A$:$B$ where $A$ cites $B$, or pairs of proteins $A$:$B$ where $A$ and $B$ physically interact in the cell. My analogical reasoning setup is a formal measure of similarity between such relational structures. The motivation is to provide tools for exploring subpopulations of interest starting from a set of pairs $\mathbf{S}$ that is chosen by an expert. A measure of analogical similarity allows one to rank which other pairs behave in a quantitatively similar way to those pairs in $\mathbf{S}$.

For a more concrete example of application, I have recently started collaborating with Edoardo Airoldi, a post-doc in Princeton, on how to apply such methods to biological domains: in this case, can a machine propose pairs of proteins that interact in a way that is analogous to a set of examples chosen by an expert?

- Silva, R.; Airoldi, A. and Heller, K. (2006). "The role of analogies in biological data: a study in the exploratory analysis of protein-protein interactions." `http://www.gatsby.ucl.ac.uk/~rbas`

# 3 Future work

All of my recent work opens a whole new set of issues. Here I describe a few future directions on which I intend to work.

**Advances on Bayesian inference for mixed graph models**

It is time for mixed graph models to receive more attention. Part of the problem is the need for discrete mixed graph models, an area with still many open questions. Although we already have one way of approaching this problem, there are other possibilities. For instance, the parameterization of Drton and Richardson (2005), for some classes of mixed graphs, implies a different family of discrete distributions. I am considering

developing priors and algorithms for Bayesian inference within this family. Other issues include learning Markov equivalence classes of mixed graphs (which requires efficient approximations for the marginal likelihood of such models) and investigations on alternative ways of parameterizating such models. What could be considered an adequate way of expressing knowledge about unmeasured confounding in observational studies (Rosenbaum, 2002)?

**Analogical similarity in complex structures**

I am excited with the possibility of doing more extensive work in the analysis of biological data. Another issue with the current approach is its high computational cost. This can a potential problem when modeling relational structures composed of more than pairs of objects. Moreover, there is clearly a link between causal and analogical reasoning, as recently illustrated by Kemp et al. (2006). Which types of applications could explore analogical similarity between causal relations?

**The dynamic structure of unstructured data**

In our original work on workflow modeling, we raised the possibility of pulling together different unstructured (text) data sources for generating a workflow structure of communication and problem solving within an organization. This is a very ambitious goal, but special cases of this problem can be treated up to some level. One particular problem, that of tracing the evolution of topics over time (Blei and Lafferty, 2006), could be analysed under the viewpoint where the evolution of a topic might diverge on parallel threads, and such threads sometimes converge (such as the evolution of papers on different areas whose topics end up being unified at some point). This contains elements of both workflow modeling and text analysis.

# References


D. Bartholomew and M. Knott. *Latent Variable Models and Factor Analysis*. Arnold Publishers, 1999.

D. Blei and J. Lafferty. Dynamic topic models. *Proceedings of the 23rd ICML*, 2006.

K. Bollen. *Structural Equation Models with Latent Variables*. John Wiley & Sons, 1989.

M. Drton and T. Richardson. Binary models for marginal independence. *Department of Statistics, University of Washington, Tech. report 474*, 2005.

D. Dunson, J. Palomo, and K. Bollen. Bayesian structural equation modeling. *Statistical and Applied Mathematical Sciences Institute, Technical Report #2005-5*, 2005.

M. Jordan. *Learning in Graphical Models*. MIT Press, 1998.

C. Kemp, P. Shafto, A. Berke, and J. Tenenbaum. Combining causal and similarity-based reasoning. *NIPS*, 2006.

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

R. Reyment and K. Joreskog. *Applied Factor Analysis in the Natural Sciences*. Cambride University Press, 1996.

T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian J. of Statistics*, 30:145–157, 2003.

T. Richardson. Chain graphs and symmetric associations. *Learning in Graphical Models*, pages 231–259, 1998.

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

P. Rosenbaum. *Observational Studies*. Springer-Verlag, 2002.

R. Scheines, R. Hoijtink, and A. Boomsma. Bayesian estimation and testing of structural equation models. *Psychometrika*, 64:37–52, 1999.

B. Shipley. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, 2002.

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

S. Wright. Correlation and causation. *Journal of Agricultural Research*, pages 557–585, 1921.