Work-In-Progress: Ensemble Causal Learning for Modeling Post-Partum Depression

Nandini Ramanan
University of Texas, Dallas
Nandini.Ramanan@utdallas.edu

Sriraam Natarajan
University of Texas, Dallas
Sriraam.Natarajan@utdallas.edu

Abstract
In this work-in-progress paper, we speculate a method for learning causal models directly from data without any interventions or inductive bias. Our ensemble approach uncovers some interesting relations for understanding post-partum depression based on family and socio-economic factors.

Introduction
We consider the problem of full model learning of causal models from data specifically in the context of predicting post-partum depression (PPD) from data. A common argument is that when learning only from data, learning causal models is only as informative as learning a correlation model (an opaque one such as deep belief network). In this work-in-progress, we put this hypothesis to test. We aim to learn a causal model using an ensemble of models and methods. This is particularly important as scaling causal learning to large problems without interventions or bias is a significantly challenging task.

Specifically, we leverage the recent success in gradient boosting to learn dependency networks (DN) (Natarajan et al. 2012; Heckerman et al. 2000). Recall that a DN is a probabilistic graphical model that approximates the joint distribution using a product of conditionals. Hence, compared to a Bayesian Network (BN) these are uninterpretable and more importantly, approximate. However, their key advantage is that since they are products of conditionals, the conditionals can be learned in parallel and can be scaled to very large data sets.

To scale causal model learning, we first learn a DN. Then we identify and remove cycles from this DN. We consider several different metrics employed in causal models to score and remove the edges. Finally, we construct a model based on the edges that are commonly present across all the metrics (i.e., the intersection of the edges from the different methods). Contrary to popular intuition, we employ two levels of ensemble learning to uncover a causal model - first is on learning a DN using boosting and the second is on learning a causal model from several different metrics. Our evaluations on the real survey data for predicting PPD demonstrates the utility of such an approach. While we present quantitative metrics, qualitatively, the edges that are learned in this model uncover interesting findings.

Background and Related work
We briefly revisit Bayesian networks and dependency networks, and functional Gradient Boosting (FGB) next.

Bayesian Network: A Bayesian network (BN) is a directed acyclic graph \( G = (X, E) \) whose nodes \( X \) represent random variables and edges \( E \) represent the conditional influences among the variables. A BN encodes factorized joint representation as, \( P(X) = \prod_i P(X_i | Pa(X_i)) \), where \( Pa(X_i) \) is the parent set of the variable \( X_i \). It is well-known that full model learning of a BN is computationally intensive, as it involves repeated probabilistic inference inside parameter estimation which in turn is performed in each step of structure search (Chickering 1996). Therefore, much of the research has focused on approximate, local search algorithms that are generally broadly classified as constraint-based and score-based. Our work is inspired by and can be considered as extending constraint-based methods which have been discussed extensively in the context of causal structure discovery. Traditional constraint-based procedures such as PC algorithm learn a BN which is consistent with conditional independencies that are inferred from data (Spirtes, Glymour, and Scheines 1993; Margaritis and Thrun 2000). A key attractiveness of these approaches is that they are sound and complete given perfect (noise-free) data (Spirtes and Glymour 1991; Zhang 2008; Colombo and Maathuis 2014). However, they require searching over exponential space of possible causal structure and this prevents their adaptation to high-dimensional data sets (even 33+ variables (Silander and Myllymaki 2012)). Our approach can be seen as scaling such methods to large data sets by potentially identifying a cyclic dependency network that can then be transformed to a causal graph. Our hypotheses is that learning such a dependency network is scalable thus reducing the complexity of causality search.

Dependency Networks: Dependency Networks (DN) (Heckerman et al. 2000) approximate the joint distribution over the variables as a product of conditionals thus allowing for cycles. These conditionals can be learned...
locally, resulting in significant efficiency gains over other exact models, i.e., \( P(X) = \prod_{X \in X} P(X|Pa(X)) \), where \( Pa(X) \) indicates the parent set of the target variable \( X \). Since they are approximate (unlike standard Bayes Nets (BNs)), Gibbs sampling is typically used to recover the joint distribution; this approach is, however, very slow even in reasonably-sized domains. In summary, learning DNs is scalable and efficient, especially for larger data sets, but BNs are preferable for inference, interpretation, discovery and analysis.

**Functional Gradient Boosting**

Functional Gradient Boosting (FGB) (Friedman 2001) represents the conditional probability distribution of the target variable as a sigmoid over a non-parametric function \( \psi \), i.e.,

\[
P(X|Pa(X)) = \frac{1}{1+\exp(-\psi(X,Pa(X)))}
\]

and computes the gradients over the functional space instead of the parametric space. This allows for approximating the true gradient by computing the functional-gradient of each training example separately. This gradient for example \( X_i \) can be shown to be \( I(X_i = \text{true}) - P(X_i|\text{parents}(X_i)) \) where \( I \) is the indicator function which returns 1 for positive and 0 for negative examples. These gradients correspond to the difference between the true label and predicted probability of an example. At each step a weak regressor (typically a short tree) is fit to capture these gradients.

**Post-partum Depression Prediction**

Many new mothers experience mild to severe mood disorders following childbirth, including post-partum depression, which is a particularly severe form of depression. Symptoms characterizing PPD can include sadness, anxiety, irritability, fatigue, reduced libido and possibly significant behavioral changes (Beck 2001). Untreated PPD can have a significant impact on the health of both mother and infant. Furthermore, PPD’s adverse effects can diminish parenting abilities, which in turn can have a profound impact on the development of the infant. Clinical diagnosis of PPD has remained a challenging problem, additionally complicated by the fact that a high percentage of women with PPD either do not report symptoms or seek help.

In a recent post-partum depression study, demographic and other non-clinical data were collected by focusing on risk factors for early detection of PPD [citation withheld for blind review]. Participants were recruited from Facebook and Twitter. The survey itself was based on the Postpartum Depression Predictors Inventory (PDPI-R), a self-administered questionnaire, having 43 questions to collect risk factors that included demographic and psychosocial questions. Out of 173 new mothers, 25% were diagnosed with post-partum depression. Given the responses to the survey questionnaire, we previously evaluated if the questions were a high predictor of occurrence of PPD. While the performance of the learning algorithm indicated that PPD can be diagnosed effectively from survey questions, the approach did not yield insights into the interactions/influences between the questions themselves.

**Ensemble Learning of Causal Models**

Given the background on the survey data set, we now present our learning algorithm. We use bold capital letters to denote sets (e.g., \( X \)) and plain capital letters to denote set members (e.g., \( X_i \)). Note that both the risk factors (43 from survey questions) and the underlying medical condition (the target) form the set of variables. Our goal is to learn a causal model over these 44 variables.

We present a high-level overview of our framework in Figure 1. After preprocessing, we learn a DN by learning each conditional distribution using FGB (Natarajan et al. 2012). In parallel, we run three different scoring metrics to compute independence scores between variables. We remove the edges from the DN that are suggested to be removed by one of these metrics. After this step, we remove edges between independent nodes from our original DN. Resultant model is acyclic. For this new unified model, we learn the parameters and compute log-likelihood score on training data for evaluation.

Our *Ensemble Causal Learning (ECL)* algorithm has three steps: learn a DN, compute test for conditional independence between nodes and then remove the edges between the nodes that are rendered mutually independent by all the tests. The overall intuition behind this approach is fairly simple: use a scalable algorithm to handle the large number of variables and learn a dense model quickly. Since it is a potentially uninterpretable cyclic model, we remove edges based on metrics typically used in causal learning (Spirtes, Glymour, and Scheines 1993).

- **Learn a DN.** DNs allow scaling the learning task to large data. Specifically, we take an efficient approach based on the observation that trees can be used inside probabilistic models to capture context-specific independence (CSI) (Boutilier et al. 1996). To this effect, we iterate through every variable and run FGB which results in several small trees for each variable. This tree captures the (conditional) probability of that particular variable \( V_i \) conditioned on all the other variables in the model, \( V \backslash V_i \). This conditional probability table can be compactly expressed as \( P(V_i \mid V \backslash V_i) \). The advantage of this approach is that it learns the qualitative relationships (structure) and quantitative influences (parameters) simultaneously and all the conditionals in parallel. The structure is simply the set of all the variables appearing in the tree and the parameters are the distributions at the leaves.

- **Test for Conditional Independence.** Next, the goal is to convert the DN learned in the previous step to a more interpretable and potentially a causal model. This necessitates removal of cycles. We apply a series of conditional independence tests in order to learn causal structure from observational data. Specifically, We compute two statistical tests 1). Pearson’s Chi-Squared test (Pearson 1992) 2). Hilbert Schmidt Independence Criterion (Gretton et al. 2005) and one Information theoretic test 3). Conditional Mutual Information (Wyner 1978) from the data. We take a conservative approach and identify the edges that are present in all the tests. In addition, we removed the edges that had low scores in any of the criteria. This allowed us
Figure 1: Flow Chart of the Proposed Model. Grey arrows denote the flow of data, blue arrows denote the flow of the model and green arrows denote the flow of additional statistics to compute Causal BN to learn an acyclic model.

- **Parameter Learning**: Once all the uninformative edges are detected and removed, we obtain the skeleton of the graphical model. To estimate its parameters, we learned a decision tree locally using only the variables in the parent set of every node which allows us to capture CSI.

**Experiments**

While in most literature, one would quantitatively evaluate a learning algorithm, our goal in this work is different. We explore the possibility of inducing a causal model from an ensemble learner. Thus, this necessitates the employment of a more qualitative approach. Nonetheless, we perform an evaluation on both these spectrums.

For the quantitative comparison, we compare ECL to two baseline constraint based methods that have been previously used to learn Causal BNs. (1) PC algorithm (denoted PC) (Bonissone et al. 1991) with Mutual Information test (Spirtes, Glymour, and Scheines 2000), and (2) Grow-Shrink (denoted GS) algorithm with Mutual Information test (Margaritis and Thrun 2000). Table 1 presents the training log-likelihood and it can be easily observed that the proposed approach is indeed comparable or marginally better than the standard approaches.

For the qualitative aspect, we present a sub-graph of the learned network in Figure 2. A few important links stand out - for instance, PPD is potentially caused by marital satisfaction, prenatal anxiety, prenatal depression and whether someone is a citizen of the US. This is interesting as there is a debate in the community about the link between prenatal anxiety, depression and PPD. Similarly, marital satisfaction is caused by infant temperament, relationship problems and if the woman is a first-time mom. These again are interesting and provide opportunities for further investigation[citation withheld for blind review].

**Discussion**

Our ECL has some salient advantages - (1) One could parallelize several steps - learning DN, computing independence scores and detecting cycles - potentially allowing for scaling learning to large problems. (2) ECL exploits CSI at two levels - when learning a DN and when computing causal influences. (3) Standard regularization techniques can be easily adapted. (4) Finally, the use of both local search and constraint-based methods inside the algorithm enables it to learn denser models than the constraint-based methods making them an attractive option for real data sets.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Methods</th>
<th>Log-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPD</td>
<td>ECL</td>
<td>-3442.568</td>
</tr>
<tr>
<td></td>
<td>PC</td>
<td>-3481.209</td>
</tr>
<tr>
<td></td>
<td>GS</td>
<td>-3648.034</td>
</tr>
</tbody>
</table>

Table 1: Table comparing training log-likelihood estimate for ECL against the baselines.
Beyond these, our key contribution is unearthing causal relationships in understanding PPD. Many of the causal links are quite interesting - a few socially interesting ones such as being married influences marital satisfaction, the temperament of infant influences marital satisfaction, unplanned pregnancy influences marital satisfaction or that PPD influences self-esteem appear. As shown in Table 2, many of these links have support from different fields of research and have been previously considered and published. As far as we are aware, this is the first work considering many of these facts in building a causal model. Consequently, these highly interesting social questions can provide directions for further research. Validating these on larger data sets, validating the causal nature of these links using interventions and domain expertise and exploring theoretically the prospect of employing several weak influences (aka ensembles) on learning a single causal model remain interesting directions for future research.

References

<table>
<thead>
<tr>
<th>Edges Picked by ECL</th>
<th>Study Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marital_satisfaction ≺ PPD</td>
<td>(Kumar and Robson 1984; O’Hara and Swain 1996; Beck 2001)</td>
</tr>
<tr>
<td>Infant_temperament ≺ Marital_satisfaction</td>
<td>(Wright, Henggeler, and Craig 1986; Fields-Olivieri, Cole, and Maggi 2017)</td>
</tr>
<tr>
<td>Child_care_stress ≺ Marital_satisfaction</td>
<td>(Doss et al. 2009; Messmer, Miller, and Yu 2012)</td>
</tr>
<tr>
<td>First_time_mom ≺ Marital_satisfaction</td>
<td>(Cox et al. 1999; Yanikkerem, Ay, and Piro 2013)</td>
</tr>
<tr>
<td>Unplanned ≺ Marital_satisfaction</td>
<td>(Belsky and Rovine 1990; Cox et al. 1999; Yanikkerem, Ay, and Piro 2013)</td>
</tr>
<tr>
<td>Depression ≺ Marital_satisfaction</td>
<td>(Haffold et al. 1999)</td>
</tr>
</tbody>
</table>

Table 2: Causal factors for postpartum depression as picked by ECL with supporting references


Pearson, K. 1992. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In Breakthroughs in Statistics. Springer. 11–28.


