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 $G = \langle \mathbf{X}, \mathbf{E} \rangle$ whose nodes $\mathbf{X}$ represent random variables and edges $\mathbf{E}$ represent the conditional influences among the variables. A BN encodes factored joint representation as, $P(\mathbf{X}) = \prod_i P(\, X_i \mid \mathbf{Pa}(X_i)\,)$, where $\mathbf{Pa}(X_i)$ 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., $\mathbf{P}(\mathbf{X}) = \prod_{X \in \mathbf{X}} \mathbf{P}(X|\mathbf{Pa}(X))$, where $\mathbf{Pa}(X)$ 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 ψ, i.e., $\mathbf{P}(X|\mathbf{Pa}(X)) = \frac{\exp \psi(X;\mathbf{Pa}(X)))}{1+\exp \psi(X;\mathbf{Pa}(X)))}$ 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 $I(X_i = true) - P(X_i|parents(X_i))$ 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., $\mathbf{X}$) and plain capital letters to denote set members (e.g., $X_i \in \mathbf{X}$). 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, $\mathbf{V} \backslash V_i$. This conditional probability table can be compactly expressed as $P(V_i \mid \mathbf{V} \setminus V_i)$. 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

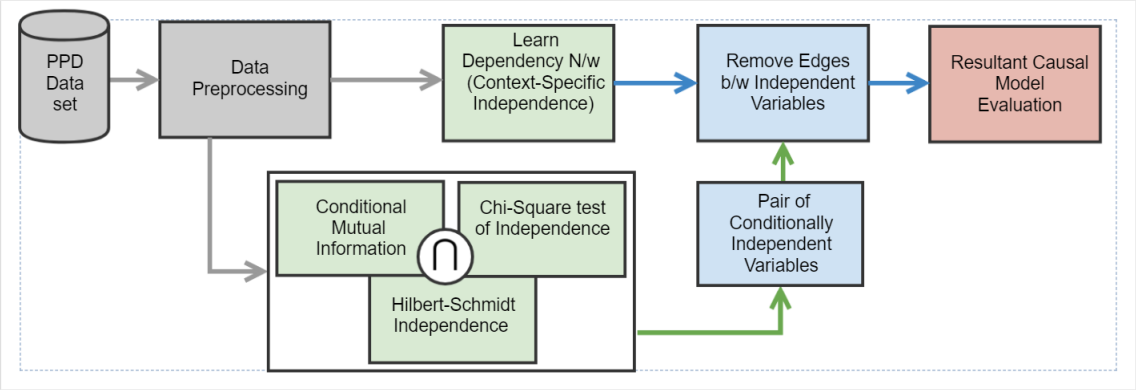

Figure 1: Flow Chart of the Proposed Model. Grey arrows denote the flow of data, blue arrows denote the flow of the model and green arrows denote the flow of additional statistics to compute Causal BN

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.

Domain	Methods	Log-Likelihood
PPD	ECL	-3442.568
	PC	-3481.209
	GS	-3648.034

Table 1: Table comparing training log-likelihood estimate for ECL against the baselines.

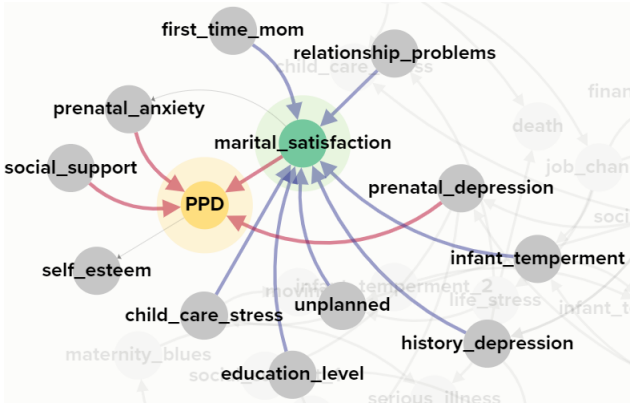

Figure 2: Subgraph from our proposed Causal BN approach depicting the most interesting influence structures. The colored edges have support in the literature as shown in Table 2. To our knowledge, none of the approaches listed in the table have considered these causal influences together and instead only considered subsets of them when performing their analyses.

Edges Picked by ECL	Study Reference
$Social\_support \prec PPD$	(O'hara and Swain 1996; Beck 2001; Seguin et al. 1999) (Nielsen et al. 2000; Logsdon, Birkimer, and Usui 2000)
$Prenatal\_anxiety \prec PPD$	(O'hara and Swain 1996; Watson et al. 1984; Beck 2001) (Johnstone et al. 2001; Neter et al. 1995; Hayworth et al. 1980)
$Prenatal\_depression \prec PPD$	(O'hara and Swain 1996; Beck 2001; Josefsson et al. 2002) (Johnstone et al. 2001; Neter et al. 1995)
$Marital\_satisfaction \prec PPD$	(Kumar and Robson 1984; O'hara and Swain 1996; Beck 2001)
$Infant\_temperament \prec Marital\_satisfaction$ $Child\_care\_stress \prec Marital\_satisfaction$	(Wright, Henggeler, and Craig 1986; Fields-Olivieri, Cole, and Maggi 2017)
$First\_time\_mom \prec Marital\_satisfaction$	(Doss et al. 2009; Messmer, Miller, and Yu 2012)
$Education \prec Marital\_satisfaction$	(Cox et al. 1999; Yanikkerem, Ay, and Piro 2013)
$Unplanned \prec Marital\_satisfaction$	(Belsky and Rovine 1990; Cox et al. 1999; Yanikkerem, Ay, and Piro 2013)
$Depression \prec Marital\_satisfaction$	(Halford et al. 1999)

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

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

Beck, C. T. 2001. Predictors of postpartum depression: an update. *Nursing research* 50(5):275–285.

Belsky, J., and Rovine, M. 1990. Patterns of marital change across the transition to parenthood: Pregnancy to three years postpartum. *Journal of Marriage and the Family* 5–19.

Bonissone, P.; Henrion, M.; Kanal, L.; and Lemmer, J. 1991. Equivalence and synthesis of causal models. In *UAI*, volume 6, 255.

Boutilier, C.; Friedman, N.; Goldszmidt, M.; and Koller, D. 1996. Context-specific independence in bayesian networks. In *UAI*, 115–123. Morgan Kaufmann Publishers Inc.

Chickering, D. 1996. Learning bayesian networks is np-complete. In *Learning from data*. Springer. 121–130.

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research* 15(1):3741–3782.

Cox, M. J.; Paley, B.; Burchinal, M.; and Payne, C. C. 1999. Marital perceptions and interactions across the transition to parenthood. *Journal of Marriage and the Family* 611–625.

Doss, B. D.; Rhoades, G. K.; Stanley, S. M.; and Markman, H. J. 2009. The effect of the transition to parenthood on relationship quality: an 8-year prospective study. *Journal of personality and social psychology* 96(3):601.

Fields-Olivieri, M. A.; Cole, P. M.; and Maggi, M. C. 2017. Toddler emotional states, temperamental traits, and their interaction: Associations with mothers and fathers parenting. *Journal of research in personality* 67:106–119.

Friedman, J. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

Halford, W. K.; Bouma, R.; Kelly, A.; and Young, R. M. 1999. Individual psychopathology and marital distress: Analyzing the association and implications for therapy. *Behavior Modification* 23(2):179–216.

Hayworth, J.; Little, B.; Carter, S. B.; Raptopoulos, P.; Priest, R.; and Sandler, M. 1980. A predictive study of post-partum depression: Some predisposing characteristics. *British Journal of Medical Psychology* 53(2):161–167.

Heckerman, D.; Chickering, D.; Meek, C.; Rounthwaite, R.; and Kadie, C. 2000. Dependency networks for inference, collaborative filtering, and data visualization. *JMLR* 1(Oct):49–75.

Johnstone, S. J.; Boyce, P. M.; Hickey, A. R.; Morris-Yates, A. D.; and Harris, M. G. 2001. Obstetric risk factors for postnatal depression in urban and rural community samples. *Australian & New Zealand Journal of Psychiatry* 35(1):69–74.

Josefsson, A.; Angelsiöö, L.; Berg, G.; Ekström, C.-M.; Gunnervik, C.; Nordin, C.; and Sydsjö, G. 2002. Obstetric, somatic, and demographic risk factors for postpartum depressive symptoms. *Obstetrics & gynecology* 99(2):223–228.

Kumar, R., and Robson, K. M. 1984. A prospective study of emotional disorders in childbearing women. *The British Journal of Psychiatry* 144(1):35–47.

Logsdon, M. C.; Birkimer, J. C.; and Usui, W. M. 2000. The link of social support and postpartum depressive symptoms in african-american women with low incomes. *MCN: The American Journal of Maternal/Child Nursing* 25(5):262–266.

Margaritis, D., and Thrun, S. 2000. Bayesian network induction via local neighborhoods. In *NIPS*, 505–511.

Messmer, R.; Miller, L. D.; and Yu, C. M. 2012. The relationship between parent-infant bed sharing and marital satisfaction for mothers of infants. *Family relations* 61(5):798–810.

Natarajan, S.; Khot, T.; Kersting, K.; Gutmann, B.; and Shavlik, J. 2012. Gradient-based boosting for statistical relational learning: The relational dependency network case. *Machine Learning* 86(1):25–56.

Neter, E.; Collins, N. L.; Lobel, M.; and Dunkel-Schetter, C. 1995. Psychosocial predictors of postpartum depressed mood in socioeconomically disadvantaged women. *Women's health (Hillsdale, NJ)* 1(1):51–75.

Nielsen, D.; Videbech, P.; Hedegaard, M.; Dalby, J.; and Secher, N. J. 2000. Postpartum depression: identification of women at risk. *BJOG: An International Journal of Obstetrics & Gynaecology* 107(10):1210–1217.

O'hara, M. W., and Swain, A. M. 1996. Rates and risk of postpartum depressiona meta-analysis. *International review of psychiatry* 8(1):37–54.

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.

Seguin, L.; Potvin, L.; St-Denis, M.; and Loiselle, J. 1999. Depressive symptoms in the late postpartum among low socioeconomic status women. *Birth* 26(3):157–163.

Silander, T., and Myllymaki, P. 2012. A simple approach for finding the globally optimal bayesian network structure. *arXiv preprint arXiv:1206.6875*.

Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review* 9(1):62–72.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, prediction, and search*. Springer.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, prediction, and search*. MIT press.

Watson, J.; Elliott, S.; Rugg, A.; and Brough, D. 1984. Psychiatric disorder in pregnancy and the first postnatal year. *The British Journal of Psychiatry* 144(5):453–462.

Wright, P. J.; Henggeler, S. W.; and Craig, L. 1986. Problems in paradise?: A longitudinal examination of the transition to parenthood. *Journal of Applied Developmental Psychology* 7(3):277–291.

Wyner, A. D. 1978. A definition of conditional mutual information for arbitrary ensembles. *Information and Control* 38(1):51–59.

Yanikkerem, E.; Ay, S.; and Piro, N. 2013. Planned and unplanned pregnancy: effects on health practice and depression during pregnancy. *Journal of Obstetrics and Gynaecology Research* 39(1):180–187.

Zhang, J. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* 172(16-17):1873–1896.