Implicit Causal Models for Genome-wide Association Studies

Dustin Tran
Columbia University

David M. Blei
Columbia University

Abstract

Progress in probabilistic generative models has accelerated, developing richer models with neural architectures, implicit densities, and with scalable algorithms for their Bayesian inference. However, there has been limited progress in models that capture causal relationships, for example, how individual genetic factors cause major human diseases. In this work, we describe *implicit causal models*, a class of causal models that leverages neural architectures with an implicit density. Further, we describe an implicit causal model that adjusts for confounders by sharing strength across examples. In experiments, we scale Bayesian inference on up to a billion genetic measurements. We achieve state of the art accuracy for identifying causal factors: we significantly outperform existing genetics methods by an absolute difference of 15-45.3%.[1]

1 Introduction

Probabilistic models provide a language for specifying rich and flexible generative processes [14, 13]. Recent advances expand this language with neural architectures, implicit densities, and with scalable algorithms for their Bayesian inference [19, 24]. However, there has been limited progress in models that capture high-dimensional causal relationships [15, 22, 8]. Unlike models which learn statistical relationships, causal models let us manipulate the generative process and make counterfactual statements, that is, what would have happened if the distributions changed.

As the running example in this work, consider genome-wide association studies (GWAS) [26, 17, 9]. The goal of GWAS is to understand how genetic factors, i.e., single nucleotide polymorphisms (SNPs), cause traits to appear in individuals. Understanding this causation both lets us predict whether an individual has a genetic predisposition to a disease and also understand how to cure the disease by targeting the individual SNPs that cause it.

We synthesize ideas from causality and modern probabilistic modeling. First, we develop *implicit causal models*, a class of causal models that leverages neural architectures with an implicit density. With GWAS, implicit causal models generalize previous methods to capture important nonlinearities, such as gene-gene and gene-population interaction. Building on this, we describe an implicit causal model that adjusts for population-confounders by sharing strength across examples (genes). In experiments, we scale Bayesian inference on implicit causal models on up to a billion genetic measurements. We achieve state of the art accuracy for identifying causal factors: we significantly outperform existing genetics methods by an absolute difference of 15-45.3%.

2 Implicit Causal Models

Probabilistic Causal Models. Probabilistic causal models [15], or structural equation models, represent variables as deterministic functions of noise and other variables. As illustration, consider the causal diagram in Figure 1. It represents a causal model where there is a global variable

$$\beta = f_\beta(\epsilon_\beta), \qquad \epsilon_\beta \sim s(\cdot),$$

[1] A longer version of this work is available at https://arxiv.org/abs/1710.10742 [23].

Workshop on Deep Learning for Physical Sciences (DLPS 2017), NIPS 2017, Long Beach, CA, USA.

Figure 1: Probabilistic causal model. **(left)** Variable x causes y coupled with a shared variable β. **(right)** A more explicit diagram where variables (denoted with a square) are a deterministic function of other variables and noise ϵ (denoted with a triangle).

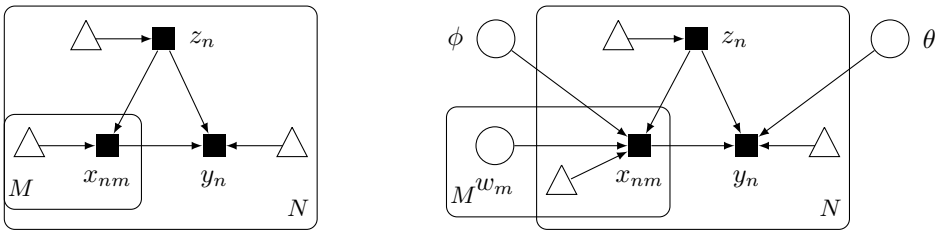

Figure 2: **(left)** Causal graph for GWAS. The population structure of SNPs for each individual (z_n) confounds inference of how each SNP (x_{nm}) causes a trait of interest (y_n). **(right)** Implicit causal model for GWAS. Its structure is the same as the causal graph but also places priors over parameters ϕ and θ and with a latent variable w_m per SNP.

and for each data point $n = 1, \dots, N$,

$$x_n = f_x(\epsilon_{x,n}, \beta), \qquad y_n = f_y(\epsilon_{y,n}, x_n, \beta), \qquad \epsilon_{x,n}, \epsilon_{y,n} \sim s(\cdot). \tag{1}$$

The noise ϵ are background variables, representing unknown external quantities which are jointly independent. Each variable β, x, y is a function of other variables and its background variable.

We are interested in estimating the causal mechanism f_y. It lets us calculate quantities such as the causal effect $p(y \mid \text{do}(X = x), \beta)$, the probability of an outcome y given that we force X to a specific value x and under fixed global structure β. This quantity differs from the conditional $p(y \mid x, \beta)$. The conditional takes the model and filters to the subpopulation where $X = x$; in general, the processes which set X to that value may also have influenced Y. Thus the conditional is not the same as if we had manipulated X directly [15].

Implicit Causal Models. Implicit models capture an unknown distribution by hypothesizing about its generative process [4, 24]. For a distribution $p(x)$ of observations x, recent advances define a function g that takes in noise $\epsilon \sim s(\cdot)$ and outputs x given parameters θ,

$$x = g(\epsilon \mid \theta), \quad \epsilon \sim s(\cdot). \tag{2}$$

Unlike models which assume additive noise, setting g to be a neural network enables multilayer, nonlinear interactions. Implicit models also separate randomness from the transformation; this imitates the structural invariance of causal models (Equation 1).

To enforce causality, we define an *implicit causal model* as a probabilistic causal model where the functions g form structural equations, that is, causal relations among variables. Implicit causal models extend implicit models in the same way that causal networks extend Bayesian networks [16] and path analysis extends regression analysis [25]. They are nonparametric structural equation models where the functional forms are themselves learned.

3 Implicit Causal Models with Latent Confounders

Consider the running example of genome-wide association studies (GWAS) (Figure 2). There are N data points (individuals). Each data point consists of an input vector of length M (measured SNPs), $x_n = [x_{n1}, \dots, x_{nM}]$ and a scalar outcome y_n (trait of interest). Typically, the # of measured SNPs M ranges from 100,000 to 1 million and the # of individuals N ranges from 500 to 10,000.

We are interested in how changes to each SNP X_m cause changes to the trait Y. Formally, this is the causal effect $p(y \mid \text{do}(x_m), x_{-m})$, which is the probability of an outcome y given that we force SNP

$X_m = x_m$ and consider fixed remaining SNPs x_{-m}. Standard inference methods are confounded by the unobserved population structure of SNPs for each individual, as well as the individual's cryptic relatedness to other samples in the data set. This confounder is represented as a latent variable z_n, which influences x_{nm} and y_n for each data index n; see Figure 2. Because we do not observe the z_n's, the causal effect $p(y \mid \text{do}(x_m), x_{-m})$ is unidentifiable [22].

Building on previous GWAS methods [17, 26, 1], we build a model that jointly captures z_n's and the mechanisms for $X_m \to Y$. Consider the implicit causal model where for each data point $n = 1, \ldots, N$ and for each SNP $m = 1, \ldots, M$,

$$z_n = g_z(\epsilon_{z_n}), \qquad x_{nm} = g_{x_m}(\epsilon_{x_{nm}}, z_n \mid w_m), \qquad y_n = g_y(\epsilon_{y_n}, x_{n,1:M}, z_n \mid \theta), \tag{3}$$

where $\epsilon_{z_n}, \epsilon_{x_{nm}}, \epsilon_{y_n} \sim s(\cdot)$. The function $g_z(\cdot)$ for the confounder is fixed. Each function $g_{x_m}(\cdot \mid w_m)$ per SNP depends on the confounder and has parameters w_m. The function $g_y(\cdot \mid \theta)$ for the trait depends on the confounder and all SNPs, and it has parameters θ. We place priors over the parameters $p(w_m)$ and $p(\theta)$.

Figure 2 (right) visualizes the model. It is a model over the full causal graph (Figure 2 (left)) and differs from the unconfounded case: § 2 only requires a model from $X \to Y$, and the rest of the graph is "ignorable" [8].

To estimate the mechanism f_y, we calculate the posterior of the outcome parameters θ,

$$p(\theta \mid \mathbf{x}, \mathbf{y}) = \int p(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) p(\theta \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) \, d\mathbf{z}. \tag{4}$$

Note how this accounts for the unobserved confounders: it assumes that $p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ accurately reflects the latent structure. In doing so, we perform inferences for $p(\theta \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$, averaged over posterior samples from $p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$.

Generative Process of Confounders z_n. We use standard normal noise and set the confounder function $g_z(\cdot)$ to the identity. This implies the distribution of confounders $p(z_n)$ is standard normal. Their dimension $z_n \in \mathbb{R}^K$ is a hyperparameter.

Generative Process of SNPs x_{nm}. Designing nonlinear processes that return matrices is an ongoing research direction (e.g., [10, 11]). To design one for GWAS (the SNP matrix), we build on an implicit modeling formulation of factor analysis; it has been successful in GWAS applications [17, 21]. Let each SNP be encoded as a 0, 1, or 2 to denote the three possible genotypes. This is unphased data, where 0 indicates two major alleles; 1 indicates one major and one minor allele; and 2 indicates two minor alleles. Set

$$\text{logit}\, \pi_{nm} = z_n^\top w_m, \qquad x_{nm} = \mathbb{I}[\epsilon_1 > \pi_{nm}] + \mathbb{I}[\epsilon_2 > \pi_{nm}], \qquad \epsilon_1, \epsilon_2 \sim \text{Uniform}(0, 1).$$

This defines a $\text{Binomial}(2, \pi_{nm})$ distribution on x_{nm}. Analogous to generalized linear models, the Binomial's logit probability is linear with respect to z_n. We then sum up two Bernoulli trials: they are represented as indicator functions of whether a uniform sample is greater than the probability. (The uniform noises are newly drawn for each index n and m.) We relax this generative process using a neural network over concatenated inputs, $\text{logit}\, \pi_{nm} = \text{NN}([z_n, w_m] \mid \phi)$. Similar to the above, the variables w_m serve as principal components. The neural network takes an input of dimension $2K$ and outputs a scalar real value; its weights and biases ϕ are shared across SNPs m and individuals n. We place a standard normal prior over ϕ.

Generative Process of Traits y_n. To specify the traits, we build on an implicit modeling formulation of linear regression. It is the mainstay tool in GWAS applications [17, 21]. Formally, for real-valued $y \in \mathbb{R}$, we model each observed trait as

$$y_n = \text{NN}([x_{n,1:M}, z_n, \epsilon] \mid \theta), \qquad \epsilon_n \sim \text{Normal}(0, 1).$$

The neural net takes an input of dimension $M + K + 1$ and outputs a scalar real value; for categorical outcomes, the output is discretized over equally spaced cutpoints. We also place a group Lasso prior on weights connecting a SNP to a hidden layer. This encourages sparse inputs: we suspect few SNPs affect the trait [27]. We use standard normal for other weights and biases.

4 Likelihood-Free Variational Inference

We described a rich causal model for how SNPs cause traits and that can adjust for latent population-confounders. Given GWAS data, we aim to infer the posterior of outcome parameters θ (Equation 4).

Trait	**ICM**	PCA (Price+06)	LMM (Kang+10)	GCAT (Song+10)
HapMap	**99.2**	34.8	30.7	**99.2**
TGP	**85.6**	2.7	43.3	70.3
HGDP	**91.8**	6.8	40.2	72.3
PSD ($a = 1$)	**97.0**	80.4	92.3	95.3
PSD ($a = 0.5$)	**94.3**	79.5	90.1	93.6
PSD ($a = 0.1$)	**92.2**	38.1	38.6	90.4
PSD ($a = 0.01$)	**92.7**	24.2	35.1	90.7
Spatial ($a = 1$)	**90.9**	56.4	60.0	75.2
Spatial ($a = 0.5$)	**86.2**	50.5	46.6	72.5
Spatial ($a = 0.1$)	**80.9**	2.4	26.6	35.6
Spatial ($a = 0.01$)	**75.5**	1.8	15.3	30.2

Table 1: Precision accuracy over an extensive set of configurations and methods; we average over 100 simulations for a grand total of 4,400 fitted models. The setting a in PSD and Spatial determines the amount of sparsity in the latent population structure: lower a means higher sparsity. ICM is significantly more robust to spurious associations, outperforming other methods by up to 45.3%.

Calculating this posterior reduces to calculating the joint posterior of confounders z_n, SNP parameters w_m and ϕ, and trait parameters θ,

$$p(z_{1:N}, w_{1:M}, \phi, \theta \mid \mathbf{x}, \mathbf{y}) \propto p(\phi)p(\theta) \prod_{n=1}^{N} [p(z_n)p(y_n \mid x_{n,1:M}, z_n, \theta) \prod_{m=1}^{M} p(w_m)p(x_{nm} \mid z_n, w_m, \phi)].$$

This means we can use typical inference algorithms on the joint posterior. We then collapse variables to obtain the marginal posterior of θ. (For Monte Carlo methods, we drop the auxiliary samples; for variational methods, it is given if the variational family follows the posterior's factorization.)

One difficulty is that with implicit models, evaluating the density is intractable: it requires integrating over a nonlinear function with respect to a high-dimensional noise (Equation 2). Thus we require likelihood-free methods, which assume that one can only sample from the model's likelihood [12, 24]. Here we apply likelihood-free variational inference (LFVI), which we scale to billions of genetic measurements [24].

5 Empirical Study: Robustness to Spurious Associations

We analyze 11 simulation configurations, where each configuration uses 100,000 SNPs and 940 to 5,000 individuals. We simulate 100 GWAS data sets per configuration for a grand total of 4,400 fitted models (4 methods of comparison). Each configuration employs a true model to generate the SNPs and traits based on real genomic data. Following Hao et al. [6], we use the Balding-Nichols model based on the HapMap dataset [2, 5]; PCA based on the 1000 Genomes Project (TGP) [3]; PCA based on the Human Genome Diversity project (HGDP) [20]; four variations of the Pritchard-Stephens-Donelly model (PSD) based on HGDP [18]; and four variations of a configuration where population structure is determined by a latent spatial position of individuals. Only 10 of the 100,000 SNPs are set to be causal.

We compare against three methods that are currently state of the art: PCA with linear regression [17] ("PCA"); a linear mixed model (EMMAX software) [9] ("LMM"); and logistic factor analysis with inverse regression [21] ("GCAT"). We use Adam with a initial step-size of 0.005, initialize neural network parameters uniformly with He variance scaling [7], and specify the neural networks for traits and SNPs as fully connected with two hidden layers, ReLU activation, and batch normalization. For the trait model's neural network, we found that including latent variables as input to the final output layer improves information flow in the network.

Table 1 displays the precision for predicting causal factors across methods. Our method achieves state of the art across all configurations. When failing to account for populations, "spurious associations" occur between genetic markers and the trait of interest, despite the fact that there is no biological connection. Precision measures a method's robustness to spurious associations: higher precision means fewer false positives and thus more robustness. Our method dominates in difficult tasks with sparse (small a), spatial (Spatial), and/or mixed membership structure (PSD): there is over a 15% margin in difference to the second best in general, and up to a 45.3% margin on the Spatial ($a = 0.01$) configuration.

Acknowledgements. DT is supported by a Google Ph.D. Fellowship in Machine Learning and an Adobe Research Fellowship. DMB is supported by NSF IIS-0745520, IIS-1247664, IIS-1009542, ONR N00014-11-1-0651, DARPA FA8750-14-2-0009, N66001-15-C-4032 , Facebook, Adobe, Amazon, and the John Templeton Foundation.

References

[1] Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471.

[2] Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. In *Human identification: The use of DNA markers*, pages 3–12.

[3] Consortium, . G. P. et al. (2010). A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061.

[4] Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series B*.

[5] Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international hapmap project.

[6] Hao, W., Song, M., and Storey, J. D. (2016). Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*.

[7] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision*. IEEE.

[8] Imbens, G. and Rubin, D. B. (2015). *Causal Inference*. Cambridge University Press.

[9] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354.

[10] Lawrence, N. D. (2005). Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *The Journal of Machine Learning Research*, 6:1783–1816.

[11] Lloyd, J. R., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Neural Information Processing Systems*.

[12] Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

[13] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.

[14] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

[15] Pearl, J. (2000). *Causality*. Cambridge University Press.

[16] Pearl, J. and Verma, T. S. (1991). A theory of inferred causation. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*.

[17] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

[18] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

[19] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.

[20] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

[21] Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature*, 47(5):550–554.

[22] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction, and search*. Springer-Verlag.

[23] Tran, D. and Blei, D. M. (2017). Implicit Causal Models for Genome-wide Association Studies. *arXiv preprint arXiv:1710.10742*.

[24] Tran, D., Ranganath, R., and Blei, D. M. (2017). Hierarchical implicit models and likelihood-free variational inference. In *Neural Information Processing Systems*.

[25] Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20(7):557–585.

[26] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

[27] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.