Causal Models for Genome-wide Association Studies

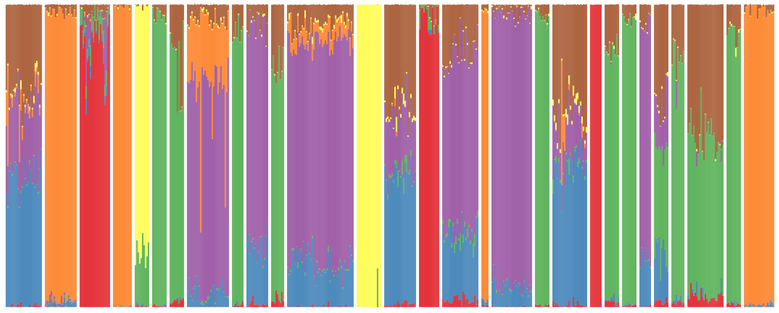
Dustin Tran[†*], David Blei[†]

[†]Columbia University, [*]Google

TL;DR

- How can we develop models to learn *causal relationships*? How can we capture latent factors which confound cause and effect?
- Using genomics as a case study, we develop causal models.
- We get SOTA, significantly outperforming baselines by 15-45.3%.

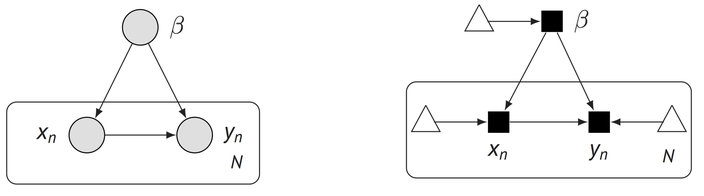
Genome-Wide Association Studies



Data consists of individuals with genetic factors x_{nm} and a trait y_n.

- Single nucleotide polymorphisms (SNPs) x_{nm} are encoded as a 0, 1, or 2. ($\approx$100K–1M)
- Phenotypes y_n may represent metabolic levels, height, disease signals. (=1)

Causal Models



Set $\beta = f_\beta(\epsilon)$. For each data point,

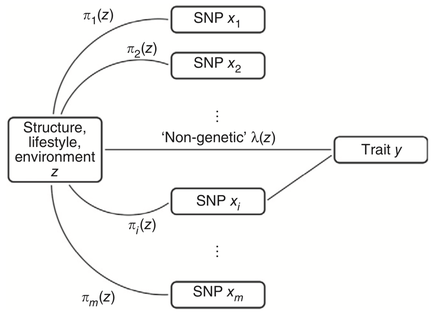
$$x_n = f_x(\epsilon, \beta), \qquad y_n = f_y(\epsilon, x_n, \beta).$$

Variables are functions of its own noise $\epsilon \sim s(\cdot)$ and other variables.

We are interested in learning the causal mechanism f_y. It lets us calculate the causal effect $p(y \mid \text{do}(X = x), \beta)$.

Under the causal graph, $p(y \mid \text{do}(x), \beta) = p(y \mid x, \beta)$. This means we can estimate f_y from observational data $\{(x_n, y_n)\}$.

Causal Model for GWAS

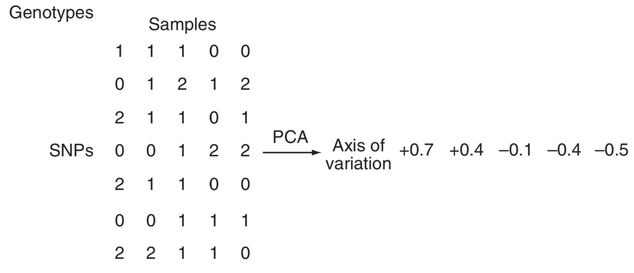


Main Idea: Build a generative model of genomes. This lets us adjust for confounders.

Posit the following causal model:

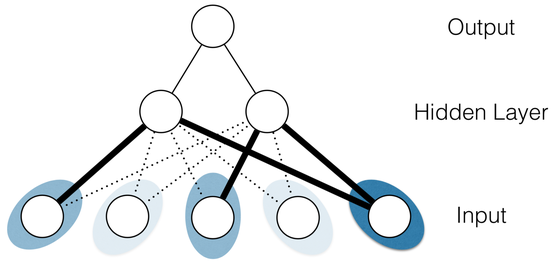
$$z = f_z(\epsilon),$$
$$x_m = f_{x_m}(\epsilon, z) \qquad \text{for each SNP } m = 1, \ldots, M,$$
$$y = f_y(\epsilon, x, z).$$

Confounders. $z_n \sim \text{Normal}(z_n; \mathbf{0}, \mathbf{I}_K)$.
It captures each person's "latent code".



SNPs. $x_{nm} \sim \text{Binomial}(2, \pi_{nm})$.
Logits are a nonlinear function of z_n and latent factors,

$$\text{logit}\, \pi_{nm} = \text{NN}([z_n, w_m] \mid \phi).$$



Traits. $y_n = \text{NN}([x_{n,1:M}, z_n, \epsilon] \mid \theta)$, $\epsilon_n \sim \text{Normal}(0, 1)$
3-layer MLP. A group Lasso prior on weights in first hidden layer encourages sparse inputs.

Causal Inference

To learn the mechanism f_y we calculate the posterior over parameters,

$$p(\theta \mid \mathbf{x}, \mathbf{y}) = \int p(\mathbf{z}, \mathbf{w}, \phi \mid \mathbf{x}, \mathbf{y}) p(\theta \mid \mathbf{x}, \mathbf{y}, \cdots)\, d\mathbf{z}\, d\mathbf{w}\, d\phi.$$

This accounts for the latent confounders: $p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$. We effectively infer the posterior of θ, averaged over samples from $p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$.

Is this principled? Our work proves $p(\theta \mid \mathbf{x}, \mathbf{y})$ provides a *consistent estimator* of the causal mechanism f_y.

How do you train it? The posterior is intractable; and the model admits an intractable likelihood. We use likelihood-free variational inference [3]. (Available in Edward!)

Semi-Synthetic Data

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

11 configurations of 100,000 SNPs and 940 to 5,000 individuals. Up to 1 billion measurements.

Implicit causal models achieve 15-45.3% higher accuracy. They are more robust to spurious associations across all experiments.

Northern Finland Birth Cohorts

Trait	**ICM**	GCAT	LMM	PCA	Uncorrected
Body mass index	0	0	0	0	0
C-reactive protein	2	2	2	2	2
Diastolic blood pressure	0	0	0	0	0
Glucose levels	3	3	2	2	2
HDL cholesterol levels	4	4	4	2	4
Height	1	1	0	0	0
Insulin levels	0	0	0	0	0
LDL cholesterol levels	3	4	3	3	3
Systolic blood pressure	0	0	0	0	0
Triglyceride levels	2	2	3	2	2

Yes. We find real-world causes.

[1] Feng, J. and Simon, N. (2017). Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*.

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

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