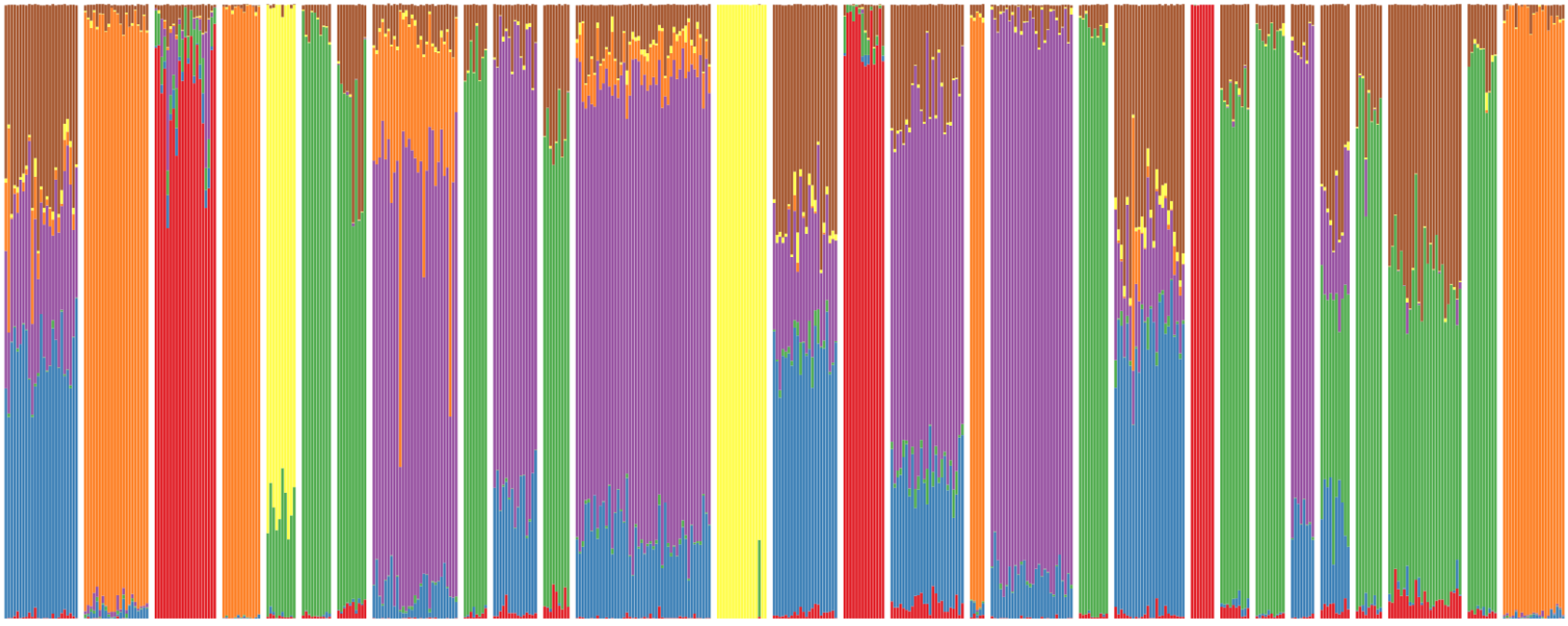
Implicit Causal Models for Genome-Wide Association Studies

Dustin Tran
Columbia University

David Blei

Genome-Wide Association Studies



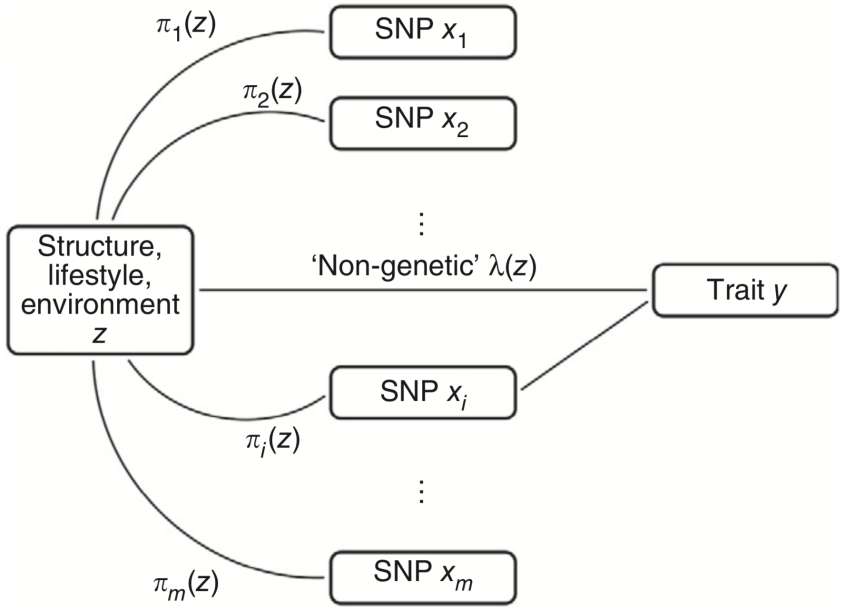
Data consists of individuals with genetic factors x_{nm} and a trait y_n.

- 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)

The goal is to understand how genetic factors cause traits in individuals.

[fig from Gopalan+ 2017]

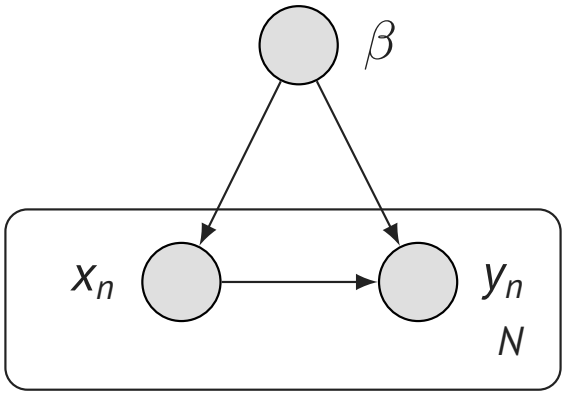
Problems in GWAS

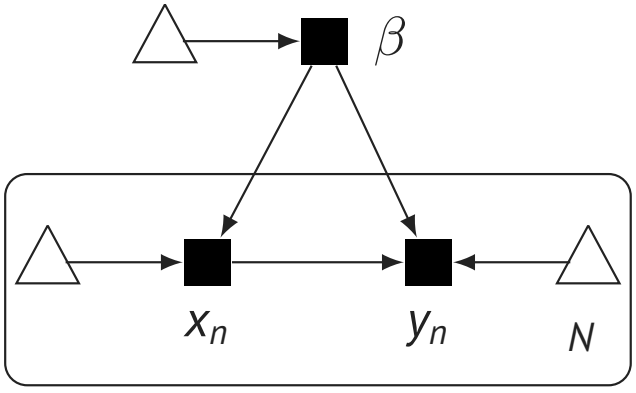


1. **Richer causal models.** Existing models apply few-to-no nonlinearities, h and engineer interactions, and assume additive Gaussian noise.
2. **Latent confounders.** 1. Latent population structure—subgroups in the population with ancestry differences. 2. relatedness among individuals.

[fig from Song+ 2015]

Background: Probabilistic Causal Models





$$\beta = f_\beta(\epsilon_\beta).$$

For each data point,

$$x_n = f_x(\epsilon_{x,n}, \beta)$$
$$y_n = f_y(\epsilon_{y,n}, x_n, \beta).$$

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

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

Background: Probabilistic Causal Models

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 $\{(x_n, y_n)\}$.

Example. An additive noise model posits

$$y_n = f(x_n, \beta \mid \theta) + \epsilon_n, \qquad \epsilon \sim s(\cdot).$$

f might be linear or use splines. With a prior $p(\theta)$, Bayesian inference yields

$$p(\theta \mid \mathbf{x}, \mathbf{y}, \beta) \propto p(\theta) p(\mathbf{y} \mid \mathbf{x}, \theta, \beta).$$

We can use standard approximate inference algorithms.

Implicit Causal Models

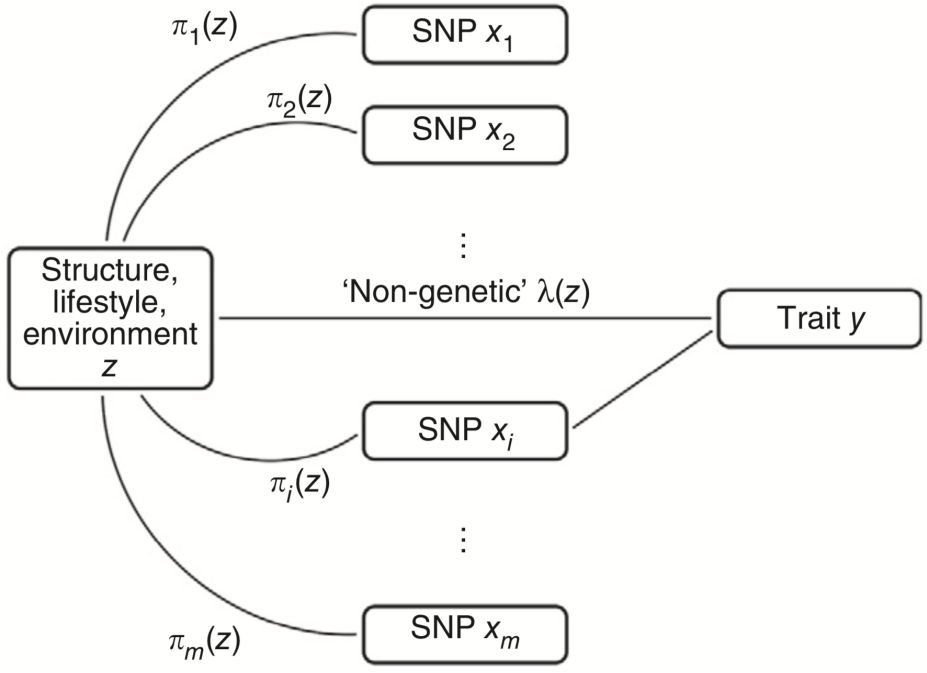
Implicit models posit a distribution via its generative process. For noise $\epsilon \sim s(\cdot)$ define a function g,

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

Setting g to a neural net enables multilayer, nonlinear interactions.

Implicit causal models are universal approximators of causal models.

Implicit Causal Models with Latent Confounder



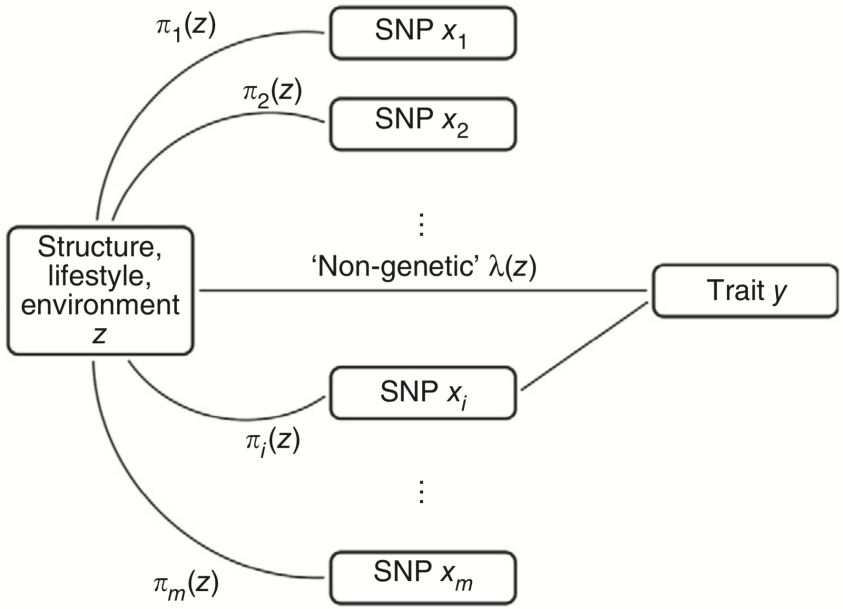
Consider a causal model for GWAS. For each SNP $m = 1, \ldots, M$,

$$
\begin{aligned}
z_n &= g_z(\epsilon_{z_n}), \\
x_{nm} &= g_{x_m}(\epsilon_{x_{nm}}, z_n \mid w_m), \\
y_n &= g_y(\epsilon_{y_n}, x_{n,1:M}, z_n \mid \theta).
\end{aligned}
$$

This is newly drawn per person n.

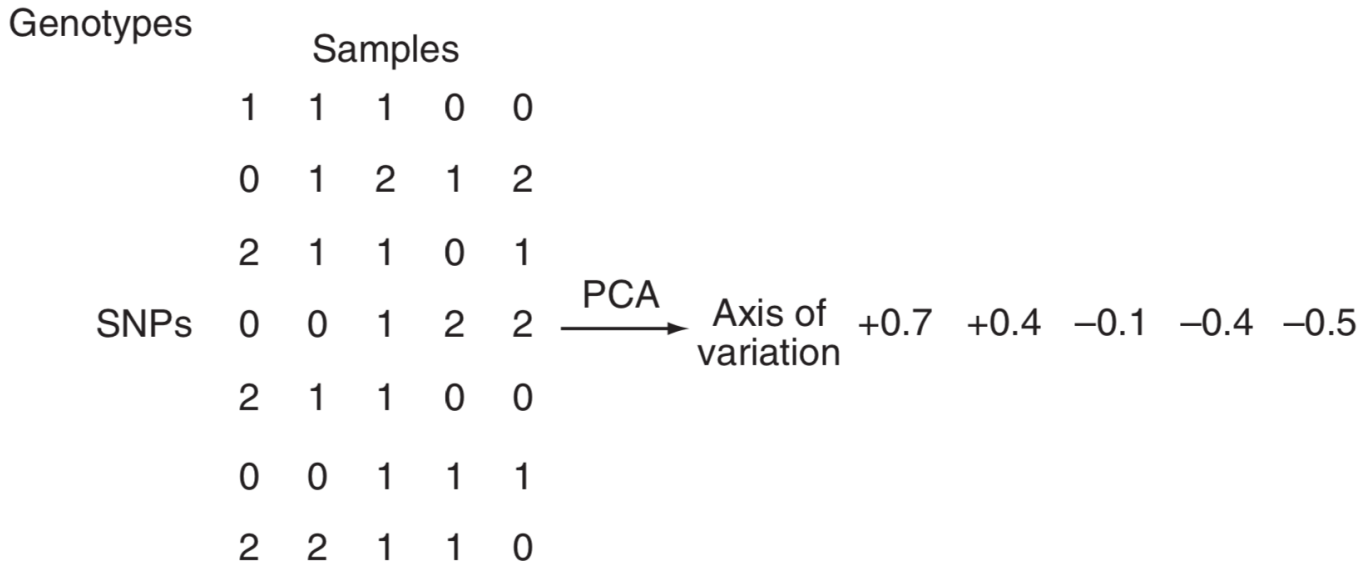
[fig from Song+ 2015]

Implicit Causal Model with a Latent Confounder



Confounders. $z_n \sim \text{Normal}(z_n; \mathbf{0}, \mathbf{I}_K)$.

Implicit Causal Model with a Latent Confounder



SNPs. $x_{nm} \sim \text{Binomial}(2, \pi_{nm})$.

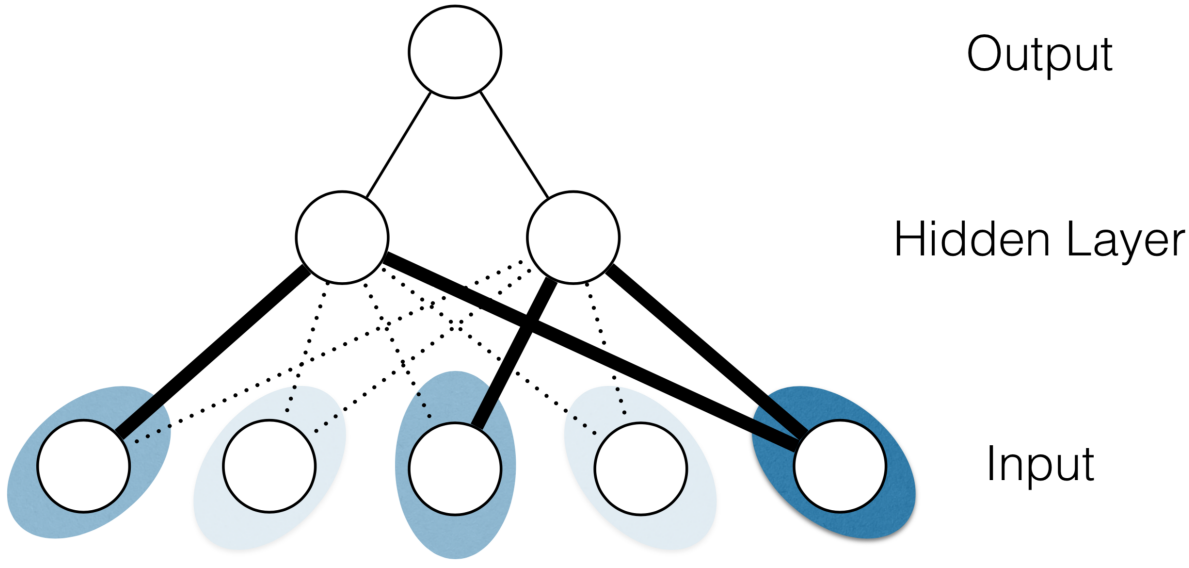
Logits are a nonlinear function of z_n and latent factors,

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

Standard normal prior over w_m and ϕ. This generalizes logistic factor analysis.

[fig from Price+ 2006]

Implicit Causal Model with a Latent Confounder



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

This generalizes linear regression.

We place a group Lasso prior on weights in first hidden layer. This encourages sparse inputs. Standard normal for others.

[fig from Feng+Simon 2017]

Causal Inference

To estimate the mechanism f_y we calculate the posterior $p(\theta \mid \mathbf{x}, \mathbf{y})$.

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

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 $p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$.

Note. Causal inference with latent confounders can be dangerous: it uses the data twice. Our work proves $p(\theta \mid \mathbf{x}, \mathbf{y})$ provides a *consistent estimator* of the causal mechanism f_y.

Causal Inference

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

The posterior is intractable. Moreover, the model admits an intractable likelihood. This bars traditional algorithms.

We use **likelihood-free variational inference**. We scale it to millions of genetic factors. (Available in Edward!)

Simulation Study

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.

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