UCLA

Tractable and expressive generative models of genetic variation data

Meihua Dang
University of California, Los Angeles
Sriram Sankararaman*
University of California, Los Angeles

Anji Liu
University of California, Los Angeles
Guy Van den Broeck*
University of California, Los Angeles

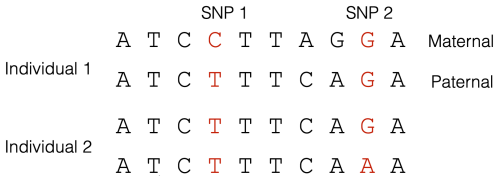
Xinzhu Wei
University of California, Los Angeles

* *Equal Contribution*

May 23th, 2022 - ***International Conference on Research in Computational Molecular Biology***

*Tractable and expressive generative models of **genetic variation data***

DNA sequence data

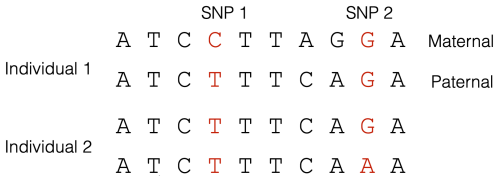


Can be represented as

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

*Tractable and expressive generative models of **genetic variation data***

DNA sequence data



Can be represented as

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

It has wide applications

$\Rightarrow$ *genotype imputation, haplotype phasing, ancestry inference*

but challenging to learn

- high-dimensional
- data-scarce
- unevenly-distributed

$\Rightarrow$ *1000 Genomes Project studies 2500 individuals, discovers 90 million SNPs, includes >99% of SNPs with with frequency > 1%*

*Tractable and expressive **generative models** of genetic variation data*

Probabilistic models that represent joint probability $p(\mathbf{X})$ over random variables $\mathbf{X}$.

- traditional probabilistic models, such as HMM and Markov chain
- more recent ML approaches, such as VAEs and GANs

Tractable and expressive *generative models of genetic variation data*

expressive

$\Rightarrow$ *how well it captures the data*

tractable

$\Rightarrow$ *ability for probabilistic inferences, such as likelihoods, MAP*

Tractable and expressive *generative models of genetic variation data*

expressive

$\Rightarrow$ *how well it captures the data*

tractable

$\Rightarrow$ *ability for probabilistic inferences, such as likelihoods, MAP*

***Q*: Can we get the benefits of both ?**

A*: *Probabilistic circuits (PCs) !

Outline

Empirical Evaluation
- Density Estimation
- Principle Component Analysis (PCA)
- Pairwise Correlation

Probabilistic Circuits (PCs)

Conclusions

Empirical Evaluation

- Density Estimation
- Principle Component Analysis (PCA)
- Pairwise Correlation

Probabilistic Circuits (PCs)

Conclusions

Density Estimation

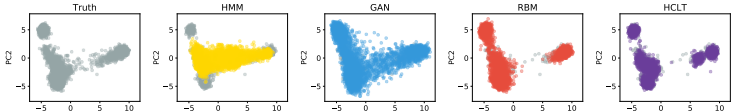
how well the models fit the data

Dataset	Indep	Markov Chain	HMM	HCLT (PC)
805	-491.10	-438.64	-402.50	**-389.20**
10K	-2390.09	-633.14	-1194.72	**-310.93**

Table: **Log-likelihoods results on subsets SNPs from 1000 Genomes Project**

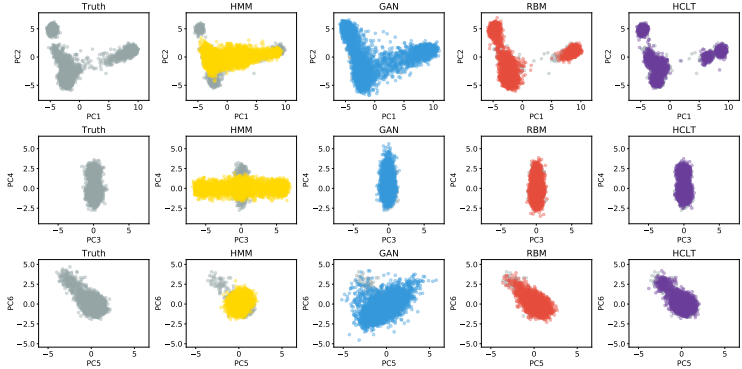
Principle Component Analysis (PCA)

how well the models fit the data

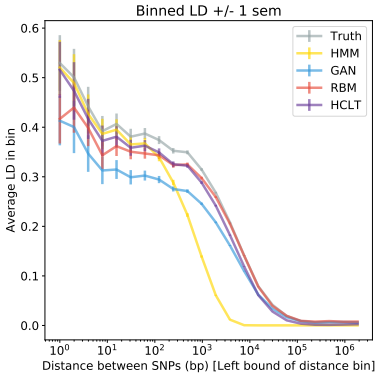


Principle Component Analysis (PCA)

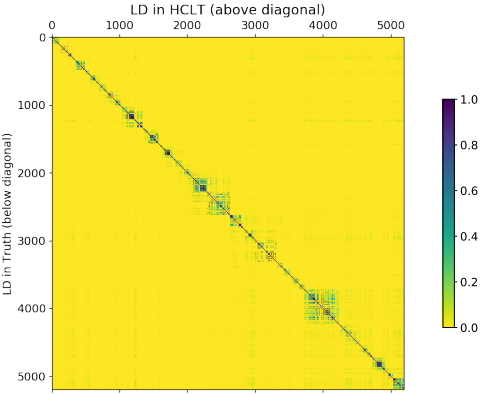
how well the models fit the data



Pairwise Correlation



(a) R^2 as a function of SNP distance



(b) R^2 matrices comparing ground truth and PC

Empirical Evaluation

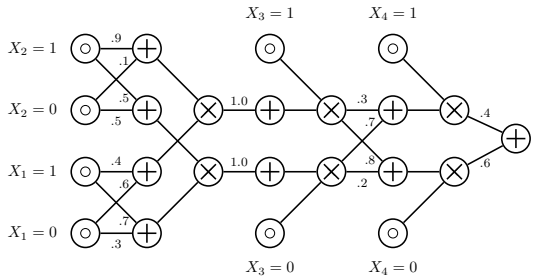
- Density Estimation
- Principle Component Analysis (PCA)
- Pairwise Correlation

Probabilistic Circuits (PCs)

Conclusions

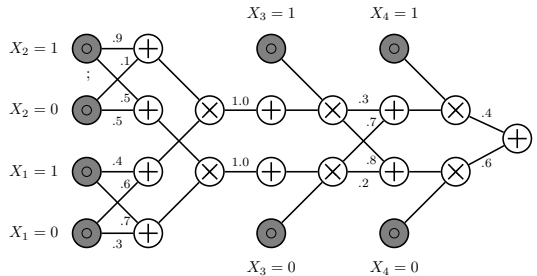
Probabilistic Circuits (semantics)

PCs encode joint distributions via computational graphs, e.g., a PC with 4 SNPs



Probabilistic Circuits (semantics)

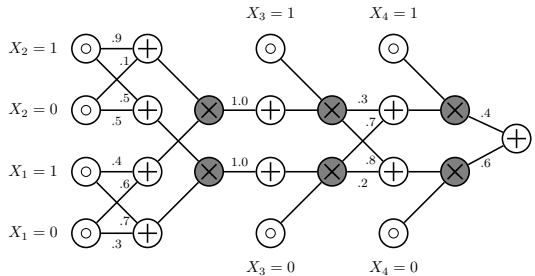
PCs encode joint distributions via computational graphs, e.g., a PC with 4 SNPs



Input nodes are tractable distributions, *e.g., indicator functions* $\mathrm{p}(X_i = 1) = [X_i = 1]$

Probabilistic Circuits (semantics)

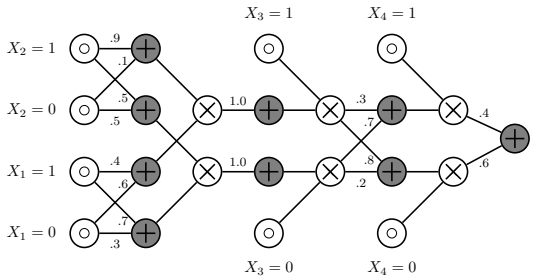
PCs encode joint distributions via computational graphs, e.g., a PC with 4 SNPs



Product nodes are factorizations $\prod_{c \in \mathsf{in}(n)} \mathrm{p}_c(\mathbf{x})$

Probabilistic Circuits (semantics)

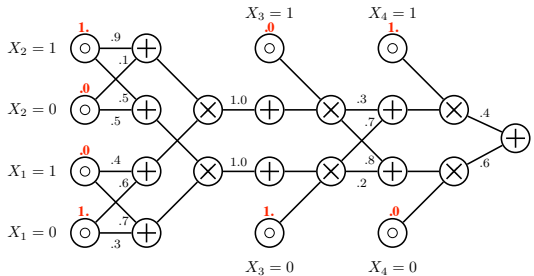
PCs encode joint distributions via computational graphs, e.g., a PC with 4 SNPs



Sum nodes are mixture models $\sum_{c \in \mathsf{in}(n)} \theta_{n,c}\, \mathrm{p}_c(\mathbf{x})$

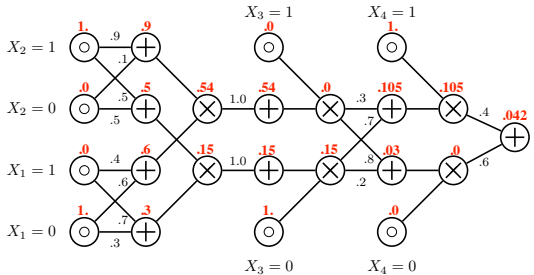
Probabilistic Circuits (tractability)

Compute likelihood $p(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 1)$



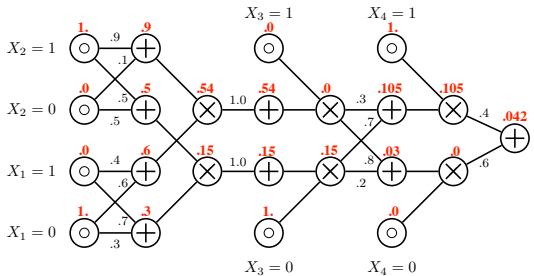
Probabilistic Circuits (tractability)

Compute likelihood $p(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 1)$



Probabilistic Circuits (tractability)

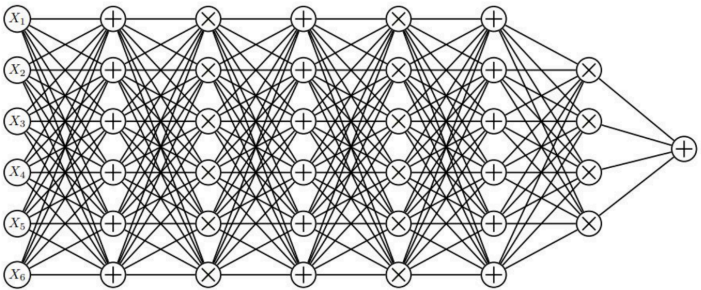
Compute likelihood $p(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 1)$



Computing likelihood is time linear in the size of PC

Probabilistic Circuits (expressiveness)

Large scale, a deep architecture, millions of parameters



Efficient learning algorithms

Conclusions

The first attempt to introduce probabilistic circuits to bio-informatics

- comparable or better performance
- tractable, expressive, time-efficient to train
- applicable to different tasks