# Non-linear Genetic Effects for Complex Traits

Yu-Wei Chen, Zhengtong Liu, Jacob Wallin

December 3, 2022

## 1 Introduction

Polygenic Risk Scores (PRS) have been a common tool used to assess the relationship between certain genotypes and phenotypes. It is especially useful for disease and therapy response evaluation, which plays a key role in personalized medicine. Over the years, literature reviews have pointed out several insufficiencies concerning PRS. Since PRS assumes a linear combination of the variables, capturing non-linearities between the variants have become a challenge; for large-scale datasets, most statistical approaches rely on algorithms that are computationally inefficient. Even though deep learning models are powerful non-linear methods for large training data and often exhibit greater predictive accuracy, they are often criticized for their limited interpretability. A recently developed model, Biologically Annotated Neural Networks (BANNs), provides the ability to compute high-dimensional genomic data as well as yield interpretable probabilistic summaries. We would like to look deeper into BANNs and compare its performance with PRS. Furthermore, we would try to extend the model by incorporating the gene-gene interaction into the model.

## 2 Biological Annotated Neuron Networks (BANNs)

### 2.1 BANN Architecture [1]

The BANNs model assumes the model:

$$\mathbf{y} = \sum_{g=i}^{G} h(\mathbf{X}_g \boldsymbol{\theta}_g + \mathbf{1} b_g^{(1)}) w_g + \mathbf{1} b^{(2)}$$

where $h(\cdot)$ is a non-linear function. It can also be written as $\mathbf{y} = \mathbf{H}(\boldsymbol{\theta})\mathbf{w} + \mathbf{1}b^{(2)}$ where $\mathbf{H}(\boldsymbol{\theta}) = [h(\mathbf{X}_g \boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g]_{\{g=1\cdots G\}}$. The Spike-and-Slab Prior is imposed to induce the sparsity of the effect sizes:

$$\theta_j \sim \pi_\theta \sum_{k=1}^{K} \eta_{\theta k} \mathcal{N}(0, \sigma_{\theta k}^2) + (1 - \pi_\theta)\delta_0$$

where $\pi_\theta$ (hyper-parameter) denotes the total proportion of SNPs that have a non-zero effect on the traits of interest; $\delta_0$ is a point mass at zero; $\boldsymbol{\eta}_\theta = (\eta_{\theta k})_{\{k=1\cdots K\}}$ represents the marginal probability that a randomly selected SNP belongs to the $k$-th mixture component (so $\sum_k \eta_{\theta k} = 1$); $\boldsymbol{\sigma}_\theta^2 = (\sigma_{\theta k}^2)_{\{k=1\cdots K\}}$ are the variance of the K nonzero mixture components. The prior on $\mathbf{w}$ is simpler:

$$w_g \sim \pi_w \mathcal{N}(0, \sigma_w^2) + (1 - \pi_w)\delta_0$$

Prior on $\pi_\theta$ and $\pi_w$: due to the lack of knowledge on the proportion of causal SNPs and SNP-sets, the prior

$$\log(\pi_\theta) \sim \mathcal{U}(-\log(J), \log(1)), \quad \log(\pi_w) \sim \mathcal{U}(-\log(G), \log(1))$$

where $J$ is the number of SNPs and $G$ is the number of SNP-sets.

## 2.2 Proposed Method

### 2.2.1 BANN is not optimal in the case of non-linear genetic effects

The key idea of BANN is to estimate the Posterior Inclusion Probability (PIP) and Phenotypic Variance Explained (PVE) through two layers of (partially) fully connected neuron network. The biological annotation is the SNP-set information, which enables the hierarchical neuron network applicable here. The sparsity-induced prior – Spike-and-Slab Prior – is imposed, and the authors use variational inference to approximate the true posterior distribution.

However, the model can be still improved when non-linear genetic effects exist. In particular, we observe that there is a significant gap between the PVE with the true heritability ($H^2$) when the pairwise interactions between SNPs contribute to the quantitative traits. Such problem is not solved when we apply the leaky ReLU non-linearity suggested in the paper. We hypothesize that the ReLU non-linearity fails to capture the genetic effects of pairwise interactions between SNPs (termed epistasis). Here we propose a modified BANN architecture.

### 2.2.2 Kernel BANN

As shown in Fig. 1, one additional layer, SNP-set (non-linear) layer is added to the architecture, while the non-linearity before the second layer (SNP-set layer) is removed. The modified model is

$$\boldsymbol{y} = \sum_{g=1}^{c}(\boldsymbol{X}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g^{(1)} + \sum_{g=1}^{c} h(\boldsymbol{X}_g)\boldsymbol{w}_g^{(2)} + \mathbf{1}b^{(2)}$$

where the first term and the third terms come from the original BANN model (the leaky ReLU function is removed), and $h(\cdot)$ in the second term denotes a kernel function (e.g., pairwise interaction). By assuming that the non-linear relationship only exists within an SNP set, the new term aims to capture such non-linear effects that are not explicitly exhibited in the original model.

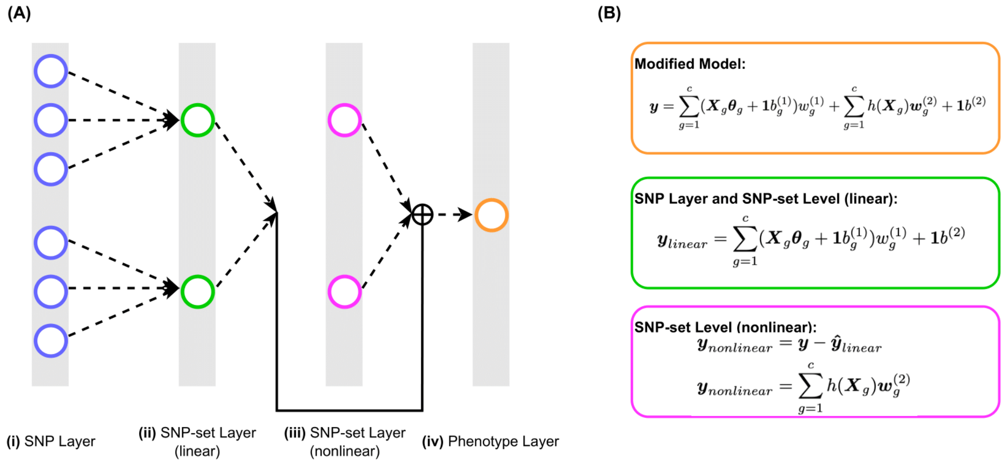
(A)
(B)
Modified Model:
$\boldsymbol{y} = \sum_{g=1}^{c}(\boldsymbol{X}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g^{(1)} + \sum_{g=1}^{c} h(\boldsymbol{X}_g)\boldsymbol{w}_g^{(2)} + \mathbf{1}b^{(2)}$
SNP Layer and SNP-set Level (linear):
$\boldsymbol{y}_{linear} = \sum_{g=1}^{c}(\boldsymbol{X}_g\boldsymbol{\theta}_g + \mathbf{1}b_g^{(1)})w_g^{(1)} + \mathbf{1}b^{(2)}$
SNP-set Level (nonlinear):
$\boldsymbol{y}_{nonlinear} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{linear}$
$\boldsymbol{y}_{nonlinear} = \sum_{g=1}^{c} h(\boldsymbol{X}_g)\boldsymbol{w}_g^{(2)}$
(i) SNP Layer
(ii) SNP-set Layer (linear)
(iii) SNP-set Layer (nonlinear)
(iv) Phenotype Layer

Figure 1: Architecture of Kernel BANN.

We implement the kernel BANN model in the following way: The first two layers are kept the same as before. We fix the linear effects, and learn the coefficients of the SNP-set layer (non-linear) from the residue of the previous two layers. The implementation details can be found at https://github.com/Zhengtong-Liu/CS-M226-Project.

## 2.3 Experiments

### 2.3.1 Simulation Design

We assume

$$\mathbf{y} = \sum_{c \in C} \mathbf{x}_c \theta_c + \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \tau^2 \mathbf{I}),$$

where $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_{c\times 1}, \boldsymbol{I}_{c\times c})$ is the additive effect size, $\mathbf{W} \sim \mathcal{N}(\mathbf{0}_{n\times e}, \boldsymbol{I}_{n\times n})$ represents the pairwise interactions between causal SNPs with corresponding effects $\boldsymbol{\varphi} \sim \mathcal{N}(\mathbf{0}_{e\times 1}, \boldsymbol{I}_{e\times e})$ ($e$ is the number of epistasis effects).

For simplicity, we assume $\mathbb{V}[\mathbf{y}] = 1$, so the broad-sense heritability $H^2 = \mathbb{V}[\sum \mathbf{x}_c \theta_c] + \mathbb{V}[\mathbf{W}\boldsymbol{\varphi}]$. The additive effect makes up $\rho$ while the pairwise interaction makes up $(1-\rho)$ of the genetic variance, so the data is rescaled such that $\mathbb{V}[\sum \mathbf{x}_c \theta_c] = \rho H^2$ and $\mathbb{V}[\mathbf{W}\boldsymbol{\varphi}] = (1-\rho)H^2$. With the constraint that $\mathbb{V}[\mathbf{y}] = 1$, the noise is rescaled so that $\mathbb{V}[\boldsymbol{\epsilon}] = 1 - H^2$.

### 2.3.2 Synthetic Genotype [1]

We first performed a series of experiments using the synthetic dataset. The minor allele frequency (MAF) for each SNP is randomly assigned to $p \in [0.05, 0.45]$, and we assume the genotype $x_{i,j} \in \{0, 1\}$ for this simulation. The synthetic genotype included 500 individuals and 1000 SNPs. The true $H^2 = 0.6$, and the results are averaged over 100 runs of the algorithm.

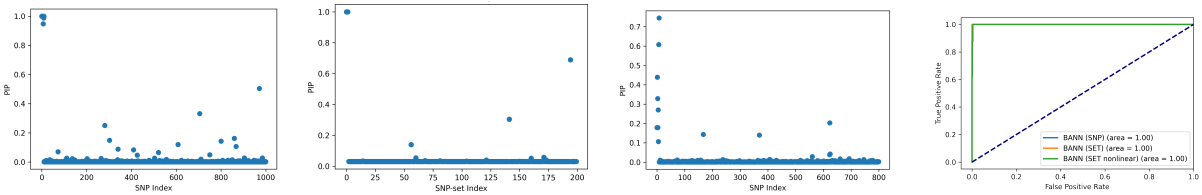

Figure 2: Posterior Inclusion Probability (PIP) plots and Receiver Operating Characteristic (ROC) curves of SNP, SET (linear and non-linear) layers for synthetic genotype data. The first 10 SNPs are causal ones. The genes that include causal genes are regarded as causal genes for the SET layers.

The results under this ideal setting are promising and serve as a sanity check for further method development. The Phenotypic Variance Explained (PVE) scores from the SNP layer, SET layer (linear), and SET layer (non-linear) are 0.48489254878656207, 0.44986504263243643, and 0.11517690230088305, with the PVE from the new layer captures the missing heritability. From Fig. 2, we see a spike on the PIP plot, and the area under the ROC curve is 1 for each layer.

### 2.3.3 Real Genotype [2]

We also used the real genotype from the 1000 Genomes Project to validate our results. We used the genotype data for a subset of 378 individuals of European ancestry from the 1000G project and approximately 15,000 SNPs on chromosome 22 from https://github.com/shz9/magenpy/tree/master/magenpy/data. For further quality control on the data, we removed SNPs with MAF of less than 5%, filtered out all variants with missing call rates, and got rid of proximal SNPs in high LD (using flags `--geno 0 --maf 0.05 --indep-pairwise 100 kb 1 0.15`). The resulting genotype has 926 SNPs. As before, the true $H^2 = 0.6$, and the results are averaged over 100 runs of the algorithm.

[1] To reproduce the results in this section, please refer to https://github.com/Zhengtong-Liu/CS-M226-Project/blob/master/BANNs/BANN/examples_docs/BANNs_example_pairwise.ipynb.

[2] To reproduce the results in this section, please refer to https://github.com/Zhengtong-Liu/CS-M226-Project/blob/master/BANNs/BANN/examples_docs/BANNs_real_geno_simulate_pheno_pairwise.ipynb.

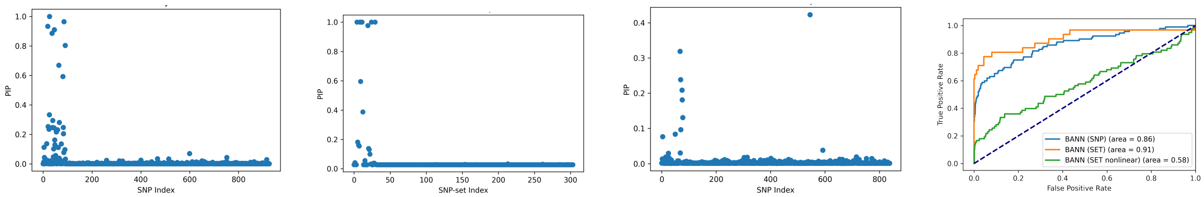

Figure 3: Posterior Inclusion Probability (PIP) plots and Receiver Operating Characteristic (ROC) curves of SNP, SET (linear and non-linear) layers for real genotype data. The first 30 SNPs are causal ones. The genes that include causal genes are regarded as causal genes for the SET layers.

The PVE scores for the SNP layer, SET layer (linear), and SET layer (non-linear) are 0.4326326467755528, 0.44188479358458965, and 0.09375916751957396. The new layer still helped to capture the non-linear effects. However, while the PIP plots show peaks at the correct positions (first 30 SNPs and corresponding genes), the ROC curves do not show very high power. Specifically, the area under the ROC curves for the SNP layer, SET layer (linear), and SET layer (non-linear) are 0.86, 0.91, and 0.58. On the one hand, it might be because some non-linear effects were regressed out or "explained away" by the first two layers, making the signal weak to capture for the new layer. On the other hand, it shows that the proposed model is not robust enough on realistic data. For future directions, we may replace the current pairwise kernel with more complex ones (e.g., RBF kernel), or we may train the third layers jointly with the previous two.

# 3 Variational Inference of Polygenic Risk Scores (VIPRS)

So far, most models developed for estimating PRS from GWAS are based on Bayesian methods because it enables a principled way to incorporate prior knowledge and provide meaningful parameters estimates. Traditionally, Bayesian PRS methods use the Markov Chain Monte Carlo (MCMC) algorithm to approximate the posterior probability, which is asymptotically accurate but can be slow to converge. This is why Bayesian approaches are restricted to a smaller subset of data (approximately one million genetic markers).
An alternative way to approximate the posterior probability is Variational Inference (VI). The approach turns posterior inference into an optimization problem, speeding up the algorithm.

## 3.1 VIPRS Architecture [2]

The VIPRS is a standard linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $X_{N\times M}$ is the genotype matrix, $\beta_{\mathbf{Mx1}}$ is a random and unknown vector of effect sizes for each of the M variants, and $\epsilon_{\mathbf{Nx1}}$ is a vector of residuals.
The setup is concerned with inferring the posterior distribution as follows:

$$\mathrm{p}(\boldsymbol{\beta}|X, y, \boldsymbol{\theta}) = \frac{\mathrm{p}(y|X, \boldsymbol{\beta}, \boldsymbol{\theta})\mathrm{p}(\boldsymbol{\beta}|\boldsymbol{\theta})}{\int \mathrm{p}(y|X, \boldsymbol{\beta}, \boldsymbol{\theta})\mathrm{p}(\boldsymbol{\beta}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\beta}}$$

where $\boldsymbol{\theta}$ is the composite term for the model hyperparameters $\pi$, $\sigma^2_{\boldsymbol{\beta}}$, $\sigma^2_{\boldsymbol{\epsilon}}$.
We also impose the spike-and-slab Prior on $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_j \sim \pi\mathcal{N}(\boldsymbol{\beta}_j; 0, \sigma^2_{\boldsymbol{\beta}}) + (1-\pi)\delta_0$$

where $\pi$ denotes the prior probability of a variant being causal, and $\sigma^2_{\boldsymbol{\beta}}$ is the prior variance on the effect size of each SNP.
For variational inference to approximate the posterior distribution of the effect sizes, propose parametric density $\mathrm{q}(\boldsymbol{\beta}, \boldsymbol{s})$ and optimize its parameters to minimize a global measure of the divergence from the true

posterior. Here we use the paired mean-field distribution family that factorizes the joint density of the effect sizes into the product of the individual densities for the effect size of each variant:

$$\mathrm{q}(\boldsymbol{\beta}, \boldsymbol{s}) = \prod_{j}^{M} q(\boldsymbol{\beta}_j, \boldsymbol{s}_j) = \prod_{j}^{M} N(\boldsymbol{\beta}_j; \boldsymbol{\mu}_j, {\sigma_j}^2) Bern(\boldsymbol{s}_j; \boldsymbol{\gamma}_j)$$

where $\boldsymbol{s}_j$ is the posterior Bernoulli variable indicating whether variant $\boldsymbol{j}$ is causal for the trait of interest.

To run VIPRS, we will need (1) GWAS summary statistics for the trait of interest and (2) Linkage-Disequilibrium (LD) matrices from an appropriately-matched reference panel. Given the unknown fixed parameters $\boldsymbol{\theta}$, the VIPRS uses the Variational Expectation-Maximization (VEM) algorithm for parameter update. In the E-Step, we update the variational parameters given the hyperparameters, while in the M-Step, we update the hyperparameters. Both steps update the free parameters with an aim of maximizing the ELBO (Evidence Lower BOund), which is used for measuring the proximity of our prediction with the true posterior.

### 3.2 Model Performance

Posterior approximation with VIPRS is computationally efficient and is able to take in as much as 10 million SNPs. Modeling with an expanded set of variants in turn significantly improves prediction accuracy for highly polygenic traits.

However, the reported performance metrics might be a lower bound due to (1) Rare variant imputation, which causes existing algorithms to elevate error rates and add substantial noise into the PRS estimate, resulting in decreased prediction accuracy. (2) Residual confounding due to population structure, which may affect effect size estimation for rare variants.

While results showed that variational approximations can be a promising alternative to MCMC, it is important to note that mean-field variational approaches are known to underestimate the posterior variances and covariances in some cases. This way, PVE tend to be underestimated, and PRS confidence intervals may also be miscalibrated. For future inplementation, we can try adopting more expressive variational families such as those derived with variational boosting, or even try combining variational methods and MCMC for more accurate results.

## 4 Theoretical Comparisons between BANNs and VIPRS

We look into two models that aim to capture genetic effects for complex traits. Since the implementation of VIPRS is not robust enough, we compare BANNs and VIPRS theoretically. Some similarities between the two models are:

- Both are built under the Bayesian framework
- Same prior assumption (spike-and-slab prior)
- Use variational inference for posterior approximation

While distinct differences include:

- BANNs has two layers, resulting in interpretable statistical outputs; VIPRS has one layer, making it computationally efficient.
- BANNs rely on gene range information; VIPRS takes in GWAS summary statistics and LD matrices.

We also observed the following characteristics of each model:

- For BANNs, the model exhibits worse performance on sparse data (i.e., a small number of SNPs).
- For VIPRS, we can perform grid search and train-test split on data, allowing finer tuning of the model parameters.

# References

[1] Pinar Demetci, Wei Cheng, Gregory Darnell, Xiang Zhou, Sohini Ramachandran, and Lorin Crawford. Multi-scale inference of genetic trait architecture using biologically annotated neural networks. *PLOS Genetics*, 17(8):1–53, 08 2021.

[2] Shadi Zabad, Simon Gravel, and Yue Li. Fast and accurate bayesian polygenic risk modeling with variational inference. *bioRxiv*, 2022.