Novel Methods for Epistasis Detection in Genome-Wide Association Studies

Lotfi Slim, Clement Chatelain, Chloé-Agathe Azencott, Jean-Philippe Vert

Novel Methods for Epistasis Detection in Genome-Wide Association Studies

Lotfi Slim[1,2,*], Clément Chatelain[2], Chloé-Agathe Azencott[1,3] and Jean-Philippe Vert[1,4]

[1] MINES ParisTech, PSL Research University,

CBIO - Centre for Computational Biology, F-75006 Paris, France,

[2] Translational Sciences, SANOFI R&D, France,

[3] Institut Curie, PSL Research University, INSERM, U900, F-75005 Paris, France,

[4] Google Brain, F-75009 Paris, France.

Abstract

As the size of genome-wide association studies (GWAS) increases, detecting interactions among single nucleotide polymorphisms (SNP) or genes associated to particular phenotypes is garnering more and more interest as a means to decipher the full genetic basis of complex diseases. Systematically testing interactions is however challenging both from a computational and from a statistical point of view, given the large number of possible interactions to consider. In this paper we propose a framework to identify pairwise interactions with a particular target variant, using a penalized regression approach. Narrowing the scope of interaction identification around a predetermined target provides increased statistical power and better interpretability, as well as computational scalability. We compare our new methods to state-of-the-art techniques for epistasis detection on simulated and real data, and demonstrate the benefits of our framework to identify pairwise interactions in several experimental settings.

1 Introduction

The amount of data generated by genome-wide association studies (GWAS) has dramatically increased in the last few years. More diseases are now being tackled with larger cohorts. Nevertheless, despite this tangible progress, our understanding of complex diseases is still limited. The classical approach in GWAS is the marginal testing for association of the phenotype of interest with each single nucleotide polymorphism (SNP) while correcting for multiple hypothesis testing. However, this fails to explain most of the phenotypic variance known to be inheritable, a phenomenon also

*Contact: lotfi.slim@mines-paristech.fr

known as missing heritability. Epigenetics and rare variants with small to moderate effects are among the reasons advanced to explain the limitations of GWAS[1,2]. In addition, high-order epistatic effects, one of the main hypotheses behind missing heritability[3], are not taken into account in marginal testing.

By constructing additive models of significant SNPs, only a small fraction of the missing heritability, as measured by narrow-sense heritability[3], is explained. For instance, the explained heritability for type II diabetes stands at 6%[4]. For height, an extensively-studied trait, the explained proportion is only 5%[5]. By revealing genetic interactions, epistasis can give an insight into the complex mapping between genotype and phenotype that cannot be extracted from marginal association testing. For instance, several epistatic mechanisms have been highlighted in the onset of Alzheimer disease[6]. Most notably, the interaction between the two genes BACE1 and APOE4 was found to be significant on four distinct datasets.

Epistasis can be defined from two different angles: biological epistasis and statistical epistasis. The definition of statistical epistasis dates back to Fisher[7] who characterizes it as the departure from additivity in a mathematical model relating multilocus genotypes to phenotypic variation. A number of strategies deployed in the context of statistical epistasis are reviewed in Cordell[8] and Niel et al.[9]. The strategies can be partitioned into two main categories: gene-gene interactions and SNP-SNP interactions. Approaching epistasis from the angle of gene-gene interactions is consistent with the definition of biological epistasis[10] as biomolecular or protein-protein interactions. Aggregator[11] and EigenEpistasis[12] are examples of tools with gene-gene interaction statistics as final output. In particular, Aggregator combines SNP-SNP interaction statistics to construct gene-level statistics. Exhaustive SNP-SNP interaction testing is still the most popular approach. It requires to correct for multiple testing using procedures such as Bonferroni correction[13] or the Benjamini-Hochberg procedure[14]. The latter is an example of false-discovery rate (FDR) procedures which are less stringent than family-wise error rate (FWER) procedures. The Bonferroni correction is a typical FWER controlling approach. For all procedures, the correction comes at the cost of poor statistical power[15]. For second-order interactions, billions of pairs of SNP must be tested, which impacts the statistical power. The decrease in statistical power is even greater for higher-order interactions. Moreover, exhaustive testing beyond second-order interactions is still unfeasible in reasonable time[16]. For increased speed, the current state-of-the-art BOOST[17] and its GPU-derivative[18] add a preliminary screening stage that ensures the survival of significant interactions. Another fast interaction search algorithm in the high-dimensional setting is the *xyz*-algorithm[19], where the interaction problem is considered from a different perspective. Instead of assessing the dependency between the product of two variables and an outcome, the pair of interest is a first variable and the Hadamard product of the outcome and a second variable. To reduce the computational overhead, the pair is projected on a set of random vectors. On the LURIC[20] GWAS dataset, the *xyz* algorithm tested more than 10^{11} interactions while being about as fast as a two-stage LASSO[21].

In addition to exhaustive statistical testing, one can also fit exhaustive regression models with linear ("marginal") effect terms and quadratic "interaction" terms. For a better inference of the interactions, Bien et al.[22] introduced

hierNET, a LASSO with hierarchy constraints between univariate and bivariate terms. When the truth is hierarchical, hierNET outperforms exhaustive regression models. Though the hierarchy constraint is plausible for many applications, it severely limits the scalability of the method to highly-dimensional problems particularly GWAS. The scope of the current release of hierNET[22] is only hundreds of predictors.

By contrast, instead of constructing exhaustive models, we focus on expanding knowledge around predetermined loci, which we refer to as "targets" in what follows. Such targets can be drawn from the literature, experiments or top hits in previous GWAS. Exhaustive genome-scale models with all pairwise terms are often computationally intensive and suffer from low statistical power. The leverage of formerly identified SNPs is then a sensible option. A lower number of interactions has to be studied with the additional guarantee that the target affects the phenotype in question. Nonetheless, a similar partial study should account for other effects of both the target and the rest of the genotype not owed to their interaction. A failure to address this issue can bias the results. In the epistasis literature, methods with such properties are lacking. In clinical trials, similar problems are encountered where the goal is to infer the treatment response variation uniquely due to the interaction between the treatment assignment and the clinical covariates. Developed specifically for this reason, propensity score[23] techniques are a common approach to achieve that. We therefore draw on those models to propose a family of model selection methods that robustly infer second-order interactions with a fixed SNP, through the formulation of different L_1-penalized regression problems. Given the high-dimensional setting, sparsity-aware methods like LASSO are well suited for model selection in genomic applications. The first category of methods developed in this work are regression approaches where the outcome combines the phenotype, the target and propensity-like quantities. The candidate SNPs are used as covariates. We also present a weighted binary classification approach. The outcome is the target, while the phenotype is included in the sample weights with the propensity score. A by-product of our work is a new framework to estimate conditional probabilities within the genome using the semi-parametric representation of the chromosomes developed for fastPHASE[24].

In the statistical literature, the selection of causal variants is a support recovery problem. For parameterized models like the LASSO, stability selection[25] is an attractive option as a model selection procedure. It aggregates the empirical selection probabilities for each variable along the LASSO path while controlling for the family-wise error rate. The original feature importance criterion in stability selection is the maximal selection probability along the stability path. In our work, we use as a criterion the area under the stability path because it better accounts for the early stages of the stability path.

In this paper, we propose a new framework to study epistasis by only focusing on the synergies with a predetermined target. By proceeding this way, the methods developed in this work improve the recovery of interacting SNPs compared to standard methods like GBOOST or LASSO with interaction terms. We evaluate the performances of our methods against two baseline models on simulated GWAS for several types of disease models. We also conduct a case study on a real GWAS dataset for type II diabetes to demonstrate the scalability of our methods and to investigate the result

differences between them.

2 Material and Methods

2.1 Setting and notations

We model genotypes and phenotypes as a triplet of random variables (X, A, Y), where Y is a discrete (e.g., in case-control studies) or continuous phenotype, $X = (X_1, \cdots, X_p) \in \{0, 1, 2\}^p$ represents a genotype with p SNPs, and A is a $(p+1)$-th target SNP of interest. The reason why we split the $p+1$ SNPs into X and A is that our goal is to detect interactions involving A and other SNPs in X. We restrict ourselves to a binary encoding of A in $\{-1, +1\}$, which allows us for example to study both recessive and dominant phenotypes, depending on how we binarize the SNP represented in A. We also introduce a version of the binarized target SNP taking values in $\{0, 1\}$ by letting $\tilde{A} = (A+1)/2$.

The target SNP A being binary, it is always possible to decompose the genotype-phenotype relationship as

$$Y = \mu(X) + \delta(X) \cdot A + \epsilon, \tag{1}$$

where ϵ is a zero mean random variable and

$$\begin{cases} \mu(X) = \dfrac{1}{2}\left[\mathbb{E}(Y|A=+1, X) + \mathbb{E}(Y|A=-1, X)\right], \\ \delta(X) = \dfrac{1}{2}\left[\mathbb{E}(Y|A=+1, X) - \mathbb{E}(Y|A=-1, X)\right]. \end{cases} \tag{2}$$

With these notations, we see from (1) that the term $\delta(X) \cdot A$ represents the marginal effect of A as well as synergistic effects between A and all SNPs in X. In the context of genomic data, we can interpret these synergies as pure epistatic effects. Furthermore, if $\delta(X)$ is sparse in the sense that it only depends on a subset of elements of X (which we call the *support* of δ), then the SNPs in the support of δ are the ones interacting with A. In other words, searching for epistasis between A and SNPs in X amounts to searching for the support of δ.

A GWAS dataset is a set of genotype-phenotype triplets $(X_i, A_i, Y_i)_{i=1,\ldots,n}$, which we model as independently and identically distributed according to the law of (X, A, Y). To estimate the support of δ from GWAS data, we propose below several models based on sparse regression and classification. The common thread between them is the use of propensity scores, which model the linkage disequilibrium (LD) dependency between the target SNP A and the rest of the genotype X. Mathematically, the propensity score $\pi(A|X)$ corresponds to the conditional probability of A given X. The balancing through the propensity scores filters out the common effects of the SNPs within X to only retain the synergistic effects with the target A. The first family of methods we propose all fall under the modified outcome

banner. In these models, an outcome that combines the phenotype Y with the target SNP A and the propensity score $\pi(A|X)$ is fitted linearly to the genomic covariates X. We propose several variants of this approach, based on several normalizations of $\pi(A|X)$ to control for estimation errors. Our second proposition is a case-only method based on the framework of outcome weighted learning (OWL) developed by Zhao et al.[26]. In this model, which is a weighted linear regression, the outcome is the target SNP A, and the covariates are the rest of the genotype X. The phenotype and the propensity score $\pi(A|X)$ are incorporated in the sample weights $Y/\pi(A|X)$. The following subsections (Sections 2.2 and 2.3) elaborate on those methods. Section 2.4 details our approach for the estimate of the propensity score $\pi(A|X)$. Finally, Section 2.5 explains how we perform model selection through stability selection.

If not stated otherwise, the full data pipeline is written in the **R** language. We also developed **epiGWAS**, an R package implementing all the methods presented in this work. The package is directly available via CRAN. The source code can also be downloaded from the GitHub repository `https://github.com/EpiSlim/epiGWAS`.

2.2 Modified outcome regression

For a given sample, only one of the two possibilities $A = +1$ or $A = -1$ is observed, making the direct estimation of $\delta(X)$ using (2) impossible from empirical GWAS data. The propensity score $\pi(A|X)$ comes into play to circumvent this problem. By considering $\tilde{A} = (A+1)/2 \in \{0, 1\}$, we can indeed rewrite (2) as:

$$\delta(X) = \frac{1}{2}\mathbb{E}\left[Y\left(\frac{\tilde{A}}{\pi(\tilde{A}=1|X)} - \frac{1-\tilde{A}}{\pi(\tilde{A}=0|X)}\right)\middle| X\right].$$

Given an estimate of $\pi(\tilde{A}|X)$, we define the modified outcome $\tilde{Y}$ of an observation (X, A, Y) as:

$$\tilde{Y} = Y\left(\frac{\tilde{A}}{\pi(\tilde{A}=1|X)} - \frac{1-\tilde{A}}{\pi(\tilde{A}=0|X)}\right), \tag{3}$$

and re-express simply

$$\delta(X) = \frac{1}{2}\mathbb{E}\left[\tilde{Y}|X\right]. \tag{4}$$

We note that our definition of modified outcome (3) generalizes that of Tian et al.[27] where it is defined as $\tilde{Y} = Y\tilde{A}$; both definitions are equivalent in the specific situation considered by Tian et al.[27] where A and X are independent, *i.e.*, $P(\tilde{A}=1|X) = P(\tilde{A}=1)$, and furthermore $P(\tilde{A}=1) = 1/2$. Our definition (3) is valid when A and X are not independent.

Given (4), we can estimate the support of δ from GWAS data by first transforming them into genotype - modified outcome pairs $(X_i, \tilde{Y}_i)_{i=1,\ldots,n}$, and then applying a model for support recovery in sparse regression of $\tilde{Y}$ given X. For that purpose we implement a stability selection procedure explained below.

Furthermore, we propose several alternative definitions of $\tilde{Y}$, which improve numerical stability and large-sample variance by controlling for the inverse of the propensity score $\pi(A|X)$. A first alternative, which we call normalized modified outcome, separately normalizes the inverses of the propensity scores of cases and controls. It is consistent and was found in empirical studies to have a lower variance than the original modified outcome estimator[28]:

$$\frac{\tilde{Y}_i}{n} = \left(\sum_{i=1}^{n} \frac{\tilde{A}_i}{\pi(\tilde{A}_i = 1|X_i)}\right)^{-1} \frac{Y_i \tilde{A}_i}{\pi(\tilde{A}_i = 1|X_i)} - \left(\sum_{i=1}^{n} \frac{1-\tilde{A}_i}{\pi(\tilde{A}_i = 0|X_i)}\right)^{-1} \frac{Y_i(1-\tilde{A}_i)}{\pi(\tilde{A}_i = 0|X_i)}.$$

However, both estimators may suffer from numerical instability because of the inverse of the propensity score weighting. If the conditional probabilities $\hat{\pi}(A_i = 0|X_i)$ or $\hat{\pi}(A_i = 1|X_i)$ are small, the weight attributed to the sample (i) can be very large relatively to other samples. The use of the inverse of the propensity scores is well-studied in the statistical literature[28,29]. A second alternative definition of $\tilde{Y}$, which we call shifted modified outcome, simply consists in the addition of a small term $\xi = 0.1$ to obtain an upper-bound on the inverse of the propensity scores:

$$\tilde{Y}_i = Y_i \left(\frac{\tilde{A}_i}{\pi(\tilde{A}_i = 1|X_i) + \xi} - \frac{1-\tilde{A}_i}{\pi(\tilde{A}_i = 0|X_i) + \xi}\right).$$

The last approach within this category, that we call robust modified outcome, is rather similar to modified outcome and normalized modified outcome. In fact, all three of them are solutions to the following system of equations:

$$\begin{cases} \sum_{i=1}^{n} \frac{\tilde{A}_i(Y_i - \mu_1))}{\pi(\tilde{A}_i=1|X)} + \eta_1 \frac{\tilde{A}_i - \pi(\tilde{A}_i=1|X)}{\pi(\tilde{A}_i=1|X)} & = 0 \\ \sum_{i=1}^{n} \frac{(1-\tilde{A}_i)(Y_i - \mu_0)}{1-\pi(\tilde{A}_i=1|X)} - \eta_0 \frac{\tilde{A}_i - \pi(\tilde{A}_i=1|X)}{1-\pi(\tilde{A}_i=1|X)} & = 0 \end{cases},$$

where $\mu_1 = \mathbb{E}\left[\mathbb{E}\left[Y^{(1)}|X\right]\right]$ and $\mu_0 = \mathbb{E}\left[\mathbb{E}\left[Y^{(0)}|X\right]\right]$.

For all (η_0, η_1), $\hat{\mu}_1 - \hat{\mu}_0 = \sum_{i=1}^{n} \tilde{Y}_i/n$ is a consistent estimator for the average risk difference $\mathbb{E}\left[\delta(X)\right]$. Modified outcome corresponds to the case $(\eta_0, \eta_1) = (\mu_0, \mu_1)$. $(\eta_0, \eta_1) = (0, 0)$ yields the second estimator, normalized modified outcome. Robust modified outcome is the solution to the above system with the smallest large-sample variance:

$$\begin{aligned} \frac{\tilde{Y}_i}{n} = &\left[\sum_{i=1}^{n} \frac{\tilde{A}_i}{\pi(\tilde{A}_i = 1|X_i)}\left(1 - \frac{C_1}{\pi(\tilde{A}_i = 1|X_i)}\right)\right]^{-1} \left(1 - \frac{C_1}{\pi(\tilde{A}_i = 1|X_i)}\right) \frac{\tilde{A}_i Y_i}{\pi(\tilde{A}_i = 1|X_i)} \\ &- \left[\sum_{i=1}^{n} \frac{1-\tilde{A}_i}{\pi(\tilde{A}_i = 0|X_i)}\left(1 - \frac{C_0}{\pi(\tilde{A}_i = 0|X_i)}\right)\right]^{-1} \left(1 - \frac{C_0}{\pi(\tilde{A}_i = 0|X_i)}\right) \frac{(1-\tilde{A}_i)Y_i}{\pi(\tilde{A}_i = 0|X_i)}, \end{aligned}$$

where,

$$\begin{cases} C_1 = & \frac{\sum_{i=1}^{n}\left((\tilde{A}_i - \pi(\tilde{A}_i=1|X_i))/\pi(\tilde{A}_i=1|X_i)\right)}{\sum_{i=1}^{n}\left((\tilde{A}_i - \pi(\tilde{A}_i=1|X_i))/\pi(\tilde{A}_i=1|X_i)\right)^2} \\ C_0 = - & \frac{\sum_{i=1}^{n}\left((\tilde{A}_i - \pi(\tilde{A}_i=1|X_i))/\pi(\tilde{A}_i=0|X_i)\right)}{\sum_{i=1}^{n}\left((\tilde{A}_i - \pi(\tilde{A}_i=1|X_i))/\pi(\tilde{A}_i=0|X_i)\right)^2} \end{cases}.$$

We can derive the expression of Robust modified outcome by using empirical estimates of η_0^* and η_1^*, the minimizers

of the large-sample variance of $\hat{\mu}_0$ and $\hat{\mu}_1$, respectively. For more details, we refer the reader to Lunceford and Davidian[28].

2.3 Outcome weighted learning

Inspired by the OWL model of Zhao et al.[26] in the context of randomized clinical trials, we now propose a second formulation as a weighted binary classification problem to estimate $\delta(X)$ and its support. Like OWL, this formulation amounts mathematically to predict A from X, where errors are penalized more or less depending on Y. We assume in this section that Y takes only nonnegative values, e.g., $Y \in \{0, 1\}$ for a case-control study. To take into account the dependency between A and X, we extend the OWL definition and consider the following function:

$$d^* \in \underset{d:\{0,1,2\}^p \to \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[\frac{Y}{\pi(A|X)}\phi\left(Ad(X)\right)\right], \tag{5}$$

where ϕ is a non-increasing loss function such as the logistic loss:

$$\forall u \in \mathbb{R}, \quad \phi(u) = \log\left(1 + e^{-u}\right). \tag{6}$$

The reason to consider this formulation is that:

Lemma 1. *The solution d^* to (5)-(6) is:*

$$\forall x \in \{0, 1, 2\}^p, \quad d^*(x) = \ln \frac{\mathbb{E}\left[Y|A = +1, X = x\right]}{\mathbb{E}\left[Y|A = -1, X = x\right]}.$$

Proof. For any $x \in \{0, 1, 2\}^p$ we see from (5) that $d^*(x)$ must minimize the function $l : \mathbb{R} \to \mathbb{R}$ defined by

$$\begin{aligned}\forall u \in \mathbb{R}, \quad l(u) &= \mathbb{E}\left[\frac{Y}{\pi(A|X = x)}\phi\left(Au\right) \mid X = x\right] \\ &= \phi(u)\mathbb{E}\left[Y \,|A = 1, X = x\right] + \phi(-u)\mathbb{E}\left[Y \,|A = -1, X = x\right]\end{aligned}$$

which is minimized when $l'(u) = 0$, i.e.,

$$\frac{\mathbb{E}\left[Y \,|A = 1, X = x\right]}{\mathbb{E}\left[Y \,|A = -1, X = x\right]} = \frac{\phi'(-u)}{\phi'(u)} = e^u.$$

□

Lemma 1 clarifies how d^* is related to δ: while δ is the difference of the expected phenotype conditioned to the two alternative values of A, d^* is the log-ratio of the same two quantities. In particular both functions have the same

sign for any genotype X. Hence we propose to estimate d^* and its support, as an approximation and alternative to estimating δ and its support, in order to capture epistatic phenomena with A.

For a given sample (X, A, Y) if we define the weight $W = Y/\pi(A|X)$, we can interpret d^* in (5) as a logistic regression classifier that predicts A from X, with errors weighted by W. Hence d^* and its support can be estimated from GWAS data by standard tools for weighted logistic regression and support estimation; we implement a stability selection procedure combined with elastic net regularized logistic regression, explained below.

In the case of qualitative GWAS studies, we encode Y as 0 for controls and 1 for cases. The regression weights W of controls thus become 0, resulting in a case-only approach for epistasis detection. Tools such as PLINK[30] and INTERSNP[31] implement optional case-only analyses, which can be more powerful in practice than a joint case-control analysis[8,32,33,34]. In the case of PLINK and INTERSNP, additional hypotheses such as the independence of gene-gene frequencies are needed to ensure the validity of the statistical test. In our case, the family of weights $\{W_i = 1/\pi(A_i|X_i)\}_{i=1,\cdots,n}$ corrects for the dependency between the target A and the genotype X. We can therefore forego such hypotheses on the data. We may even argue that the controls are indirectly included in the regression model through $\pi(A|X)$. It represents the dependency pattern within the general population, and not only within cases.

2.4 Estimate of the propensity score

As mentioned above, the propensity score $\pi(A|X)$ is recurrent in clinical trials. In such a context, A is the treatment assignment and X are the clinical covariates. The outcome for clinical trials is the treatment response. We are interested in the interaction between the treatment and the covariates to understand the main drivers for treatment response. Practitioners often opt for a parametric model for the propensity score $\pi(A|X)$ e.g. a regression model:

$$\text{logit}\left(\pi(A = 1|X)\right) = \gamma^T X.$$

It is common practice to include a number of higher-order terms to model the interaction between the clinical covariates within X. The included variables are preferably either causal (related to the response) or confounding variables (related to both the response and the treatment assignment).

For single-nucleotide polymorphisms, a similar logistic regression model is also possible to model the structural dependence between the target of interest A and the rest of the genotype X. Because of the ultra high-dimensional setting and the linkage disequilibrium along the chromosomes, we opt instead for a more structure-aware model, namely a hidden Markov model[24]. The hidden states represent contiguous clusters of phased haplotypes. The emission states correspond to SNPs. Several authors[24,35,36,37] advocate this model as a more flexible representation than haploblocks[38]. Our selection of this model was also motivated by the heavy skewness of the estimated propensity score distributions towards 0 and 1 due to the severe overfitting of regression models. In Appendix A, we provide a full characterization

of this model.

The hidden Markov model representation of the genome was developed to perform imputation, and has essentially remained confined to that application. For example, the fastPHASE software[24] based on this model leads to near-perfect imputation results, with error rates typically lower than 0.01. Among other applications, this representation has been used to construct knockoff copies of SNPs[39] to control the false discovery rate in GWAS[40]. The estimate of the propensity scores $\pi(A|X)$ is a new application of this representation in the context of genome-wide association studies.

Since the structural dependence is chromosome-wise, we only retain the SNPs located on the same chromosome as the SNP A, which we denote here by X_A. Mathematically, this is equivalent to the independence of the SNPs A and X_A from the SNPs of other chromosomes.

The pathological cases $\pi(A|X_A) \approx 1$ and $\pi(A|X_A) \approx 0$ can be avoided by the removal of all SNPs within a certain distance of A. In our implementation, we first performed an adjacency-constrained hierarchical clustering of the SNPs located on the chromosome of the target A. We fixed the maximum correlation threshold at 0.5. To alleviate strong linkage disequilibrium, we then discarded the SNPs within a three-cluster window of SNP A. Such filtering is sensible since we are looking for biological interactions between functionally-distinct regions. The neighboring SNPs are not only removed for the estimation of the propensity score, but also in the regression models searching for interactions. Other alternatives do exist to control the tails of the conditional distribution $\hat{\pi}(A|X)$. A straightforward approach is to trim the upper and lower percentiles (often the 1^{st} and 99^{th} percentiles)[41,42].

After fitting the unphased genotype model using fastPHASE, the last remaining step is the application of the forward algorithm[43] to obtain an estimate of the two potential observations $(A = 1, X_A)$ and $(A = -1, X_A)$. The Bayes theorem yields the desired propensity scores $\pi(A|X_A) = \pi(A|X)$.

2.5 Support estimation

In order to estimate the support of δ in the case of modified outcome regression (4), and of d^* in the case of OWL (5), we model both functions as linear models and estimate non-zero coefficients by elastic net regression[44] combined with stability selection[25].

More precisely, given a GWAS cohort $(X_i, A_i, Y_i)_{i=1,\dots,n}$, we first define empirical risks for a candidate linear model $x \mapsto \gamma^\top x$ for δ and d^* as respectively

$$R_1(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i - \gamma^\top X_i\right)^2, \quad R_2(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\pi(A_i|X_i)}\phi(A_i\gamma^\top X_i)\,.$$

For a given regularization parameter $\lambda > 0$ and empirical risk $R = R_1$ or $R = R_2$, we then define the elastic net

estimator:

$$\hat{\gamma_\lambda} \in \underset{\gamma}{\operatorname{argmin}}\ R(\gamma) + \lambda \left[(1-s)||\gamma||_1 + s||\gamma||_2^2/2\right] ,$$

where we fix $s = 10^{-6}$ to give greater importance to the L_1-penalization. Over a grid of values Λ for the penalization parameter λ, we subsample $N = 50$ times without replacement the whole cohort. The size of the generated subsamples $I_1, \cdots, I_N$ is $\lfloor n/2 \rfloor$. Each subsample I provides a different support for $\hat{\gamma}_\lambda$, which we note $\hat{S}^\lambda(I)$. For $\lambda \in \Lambda$, the empirical frequency of the variable X_k entering the support is then given by:

$$\hat{\omega}_k^\lambda = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}(k \in \hat{S}^\lambda(I_j)).$$

In the original stability selection procedure[25], the decision rule for including the variable k in the final model is $\max_{\lambda \in \Lambda} \hat{\omega}_k^\lambda \geq t$. The parameter t is a predefined threshold. For noisy high-dimensional data, the maximal empirical frequency along the stability path $\max_{\lambda \in \Lambda} \hat{\omega}_k^\lambda$ may not be sufficiently robust. In line with the results of Haury et al.[45], we found that the area under the stability path $\int_\lambda \hat{\omega}_k^\lambda \, \mathrm{d}\lambda$ is a better criterion for model selection. The main intuition behind the better performance is the early entry of causal variables into the LASSO path.

Finally, to determine the grid Λ, we make use of the **R** package **glmnet**[46]. We generate a log-scaled grid of 200 values $(\lambda_l)_{l=1,\cdots,200}$ between $\lambda_1 = \lambda_{max}$ and $\lambda_{200} = \lambda_{max}/100$, where $\lambda_{\max}$ is the maximum λ leading to a non-zero model. To improve the inference, we only retain the first half of the path comprised between λ_1 and λ_{100}. The benefit of a thresholded regularization path is to discard a large number of irrelevant covariates that enter the support for low values of λ.

3 Results

3.1 Simulations

Disease model

We simulate phenotypes using a logit model with the following structure:

$$\text{logit}(P(Y = 1|\tilde{A} = i, X)) = \beta_{i,V}^T X_V + \beta_W^T X_W + X_{Z_1}^T \, \text{diag}\left(\beta_{Z_1,Z_2}\right) X_{Z_2},$$

where V, W, Z_1 and Z_2 are random subsets of $\{1, \cdots, p\}$. The variables within the vector X_V interact with A. In the disease model, we also included two other sets of variables X_W and (X_{Z_1}, X_{Z_2}). The variable X_W corresponds to marginal effects while the two other variables X_{Z_1} and X_{Z_2} correspond to quadratic effects. The effect sizes $\beta_{0,V}, \beta_{1,V}, \beta_W$ and β_{Z1,Z_2} are sampled from $\mathcal{N}(0, 1)$. Given the symmetry around 0 of the effect size distributions,

the simulated cohorts are approximately equally balanced between cases and controls.

To account for the diversity of effect types in disease models, we simulate four scenarios with different overlap configurations between X_V and (X_W, X_{Z_1}). For each of the scenarios detailed below, we conducted 125 simulations: 5 sets of causal SNPs $\{A, V, W, Z_1, Z_2\} \times 5$ sets of size effects $\{\beta_{0,V}, \beta_{1,V}, \beta_W, \beta_{Z1,Z_2}\} \times 5$ replicates.

- Synergistic only effects, $|V \cap W| = 0, |V \cap Z_1| = 0, |V| = |W| = |Z_1| = |Z_2| = 8$;
- Partial overlap between synergistic and marginal effects, $|V \cap W| = 4, |V \cap Z_1| = 0, |V| = |W| = |Z_1| = |Z_2| = 8$;
- Partial overlap between synergistic and quadratic effects, $|V \cap W| = 0, |V \cap Z_1| = 4, |V| = |W| = |Z_1| = |Z_2| = 8$;
- Partial overlap between synergistic and quadratic/marginal effects, $|V \cap W| = 2, |V \cap Z_1| = 2, |V| = |W| = |Z_1| = |Z_2| = 8$.

Because of the filtering window around the SNP A, the causal SNPs (X_V, X_W, Z_1, Z_2) were sampled outside of that window. The second constraint on the causal SNPs is a lower bound on the minor allele frequencies (MAF). We fixed that bound at 0.2. The goal is to obtain well-balanced marginal distributions for the different variants. For rare variants, it is difficult to untangle the statistical power of any method from the inherent difficulty in detecting them. The lower bound is also coherent with the common disease-common variant hypothesis[47]: the main drivers of complex/common diseases are common SNPs.

Genotype simulations

We simulated genotypes using the second release of HAPGEN[48]. The underlying model for HAPGEN is the same hidden Markov model described in Appendix A. The starting point is a reference set of population haplotypes. The accompanying haplotypes dataset is the 1000 Genomes phase 3 reference haplotypes[49]. In our simulations, we only use the European population samples. The second input to HAPGEN is a fine scale recombination map. Consequently, the simulated haplotypes/genotypes exhibit the same linkage disequilibrium structure as the original data.

In comparison to the HAPGEN-generated haplotypes, the final markers density for SNP arrays is significantly reduced. For example, the sequencing technology for the WTCCC case-control consortium[50] is the Affymetrix 500K. As its name suggests, "only" five hundred thousand positions are genotyped. As most GWAS are based on SNP array data, we only extract from the simulated genotypes the markers of the Affymetrix 500K. In the subsequent QC step, we only retain common bi-allelic SNPs defined by a MAF > 0.01. We also remove SNPs that are not in a Hardy-Weinberg equilibrium $\left(p < 10^{-6}\right)$.

For iterative simulations, HAPGEN can be time-consuming, notably for large cohorts consisting of thousands of samples. Instead, we proceed in the following way: we generate once and for all a large dataset of 20 thousand samples on the 22^{nd} chromosome. To benchmark for varying sample sizes $n \in \{500, 1000, 2000, 5\,000\}$, we iteratively sample uniformly and without replacement n-times the population of $20\,000$ individuals to create 125 case-control cohorts. On the 22^{nd} chromosome, we then select $p = 5\,000$ SNPs located between the nucleotide positions $16\,061\,016$ and $49\,449\,618$. We do not conduct any posterior pruning to avoid filtering out the true causal SNPs.

Evaluation

We benchmark our new methods against two baselines. The first method is GBOOST[17], a state-of-the-art method for epistasis detection. For all SNP pairs, it implements the log-likelihood ratio test statistic to compare the goodness of fit of two models: the full logistic regression model with main effect terms and interaction terms, and the logistic regression model with main effects only. The preliminary sure screening step to discard a number of SNPs from exhaustive pairwise testing was omitted, since we are only interested in the GBOOST score for the pairs of the form (A, X_k), where X_k is the k-th SNP. The second method, which we refer to as product LASSO, originates from the machine learning community. It was developed by Tian et al.[27] to estimate interactions between a treatment and a large number of covariates. It fits an L_1-penalized logistic regression model with $A \times X$ as covariates. The variable of interest A is symmetrically encoded as $\{-1, +1\}$. Under general assumptions, Tian et al.[27] show how this model works as a good approximation to the optimal decision rule d^* (see Section 2.3).

We visualize the results of our methods in terms of receiver-operating characteristic (ROC) curves and precision-recall (PR) curves. The ROC and PR curves are derived from the stability paths. For each SNP, the score is the area under its corresponding stability path. For ROC/PR curves, no normalization is needed to bring the scores into the $[0, 1]$ range. The labels are 1 for the SNPs interacting with the target A, and 0 otherwise. The covariates and the outcome differ between our methods. That implies a different regularization path for each method and as a result, incomparable stability paths. For better interpretability and comparability between the methods, we use the position l on the stability path grid $\Lambda = (\lambda_l)$ s.t. $\lambda_l > \lambda_{l+1}$ instead of the value of λ_l for computing the area under the curve.

In Figure 1, we provide the ROC and PR curves for the fourth scenario which corresponds to a partial overlap between synergistic and quadratic/marginal effects and for a sample size $n = 500$. Because of space constraints, all ROC/PR figures and corresponding AUC tables are listed in Appendix B. The figures represent the average ROC and PR curves of the 125 simulations in each of the four scenarios. To generate those figures, we used the **R** package **precrec**[51]. It performs nonlinear interpolation in the PR space. The AUCs were computed with same package.

Regardless of the scenario and the sample size, the areas under all ROC curves are higher than 0.5. That confirms that all of them perform better than random, yet with varying degrees of success. By contrast, the overall areas under the precision-recall curves are low. The maximum area under the precision-recall curve is 0.41, attained by Modified

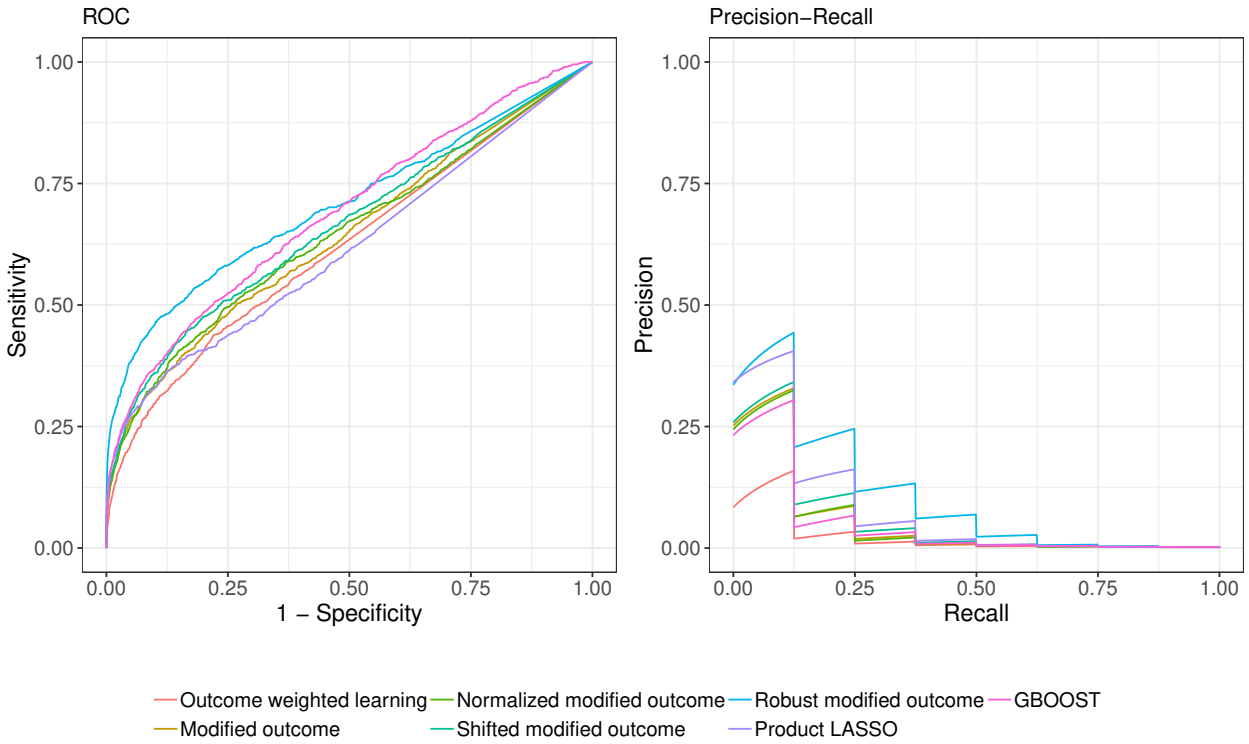

Figure 1: Average ROC (left) and PR (right) curves for the fourth scenario and $n = 500$

Outcome with shifted weights for $n = p$. This can be attributed to the imbalanced nature of the problem: 8 synergistic SNPs out of $5\,000$. For both ROC and PR, we do also observe increasing AUCs with the cohort size.

The best performing methods are robust modified outcome and GBOOST. Robust modified outcome has a slight lead in terms of ROC AUCs, notably for low sample sizes. The latter setup is the closest to our intended application in genome-wide association studies. Of special interest to us in the ROC space is the bottom-left area. It reflects the retrieval performance for highly-ranked instances. For all scenarios, we witness a better start for robust modified outcome. The other methods within the modified outcome family behave similarly. Such a result was expected because of their theoretical similarities. Despite the model misspecification, product LASSO performs rather well. On average, it comes third to GBOOST and robust modified outcome. The outcome weighted learning approach which is an approximation to estimating δ has consistently been the worst performer in the ROC space.

In PR space, the results are more mixed. For low sample sizes, robust modified outcome is still the best performing method. As the sample size increases, we observe that other methods within the modified outcome family, notably shifted modified outcome, surpass the robust modified outcome approach. Surprisingly, the good performance of GBOOST in ROC space was not reproduced in PR space. This might be explained by the highly imbalanced nature of the problem and the lower performance of GBOOST, compared to robust modified outcome in the high specificity region of the ROC curves (lower left). By contrast, product LASSO is always trailing the best performer of the modified

outcome family. As for ROC curves, we are also interested in the beginning of the PR curves. For a recall rate of 0.125, the highest precision rate is near 0.5 for the first, third and fourth scenario. That implies that we detect on average one causal SNP in the first two SNPs. For the second scenario, the highest precision rate is even higher at approximately 0.68. The area under the stability path is then a robust score for model selection in the high dimensional setting.

It is worth noting the homogeneous behavior of the different methods across the four scenarios. For a given sample size, and for a given method, the ROC and PR AUCs are similar. This suggests they all successfully filtered out the common effects term $\mu(X)$ even in presence of an overlap between the causal SNPs within $\mu(X)$ and $\delta(X)$.

3.2 Case study : type II diabetes dataset of the WTCCC

As a case study, we selected the type II diabetes dataset of the WTCCC[50] to illustrate the scalability of our methods to real datasets. To the best of our knowledge, no confirmed epistatic interactions exist for type II diabetes. We propose instead to study the synergies with a particular target: $rs41475248$ on chromosome 8. The first criterion to our choice is the presence of a significant epistatic effect. For that purpose, we initially ran GBOOST. SNP $rs41475248$ is involved in 3 epistatic interactions, when controlling for a false discovery rate of 0.05. The second criterion is being a common variant. The MAF of the selected target is 0.45.

Before running our methods on the WTCCC dataset, we applied the same QC procedures with the following thresholds: 0.01 for minor-allele frequencies and $p > 10^{-6}$ for the Hardy-Weinberg equilibrium. The number of remaining variants is $354\,439$ SNPs. The number of samples is $4\,897$, split between $1\,953$ cases and $2\,944$ controls.

To solve the different L_1-penalized regressions, we abandoned **glmnet** in favor of another solver, **biglasso**[52]. **glmnet** does not accept as input such ultra-high dimensional design matrices. On the other hand, **biglasso** was specifically developed for similar settings thanks to its multi-threaded implementation and utilization of memory-mapped files. Because **biglasso** does not implement sample weighting, it cannot be used to run outcome weighted learning. Moreover, this approach performed worse than the modified outcome approaches on simulated data, and we therefore excluded it from this case study.

The main difficulty for the evaluation of GWAS methods is the biological validation of the study results. We often lack evidence to correctly label each SNP as being involved or not in an epistatic interaction. Evaluating the model selection performance of the different methods on real datasets is then impossible. However, we can study the concordance between them. A common way to proceed is Kendall's tau which is a measure of rank correlation. In Table 1, we give the correlation matrix of our methods and the two baselines of Section 3.1. All elements are positive which indicates a relative agreement between the methods. Modified outcome, normalized modified outcome and shifted modified outcome have the highest correlation coefficients. Such a result was expected because of their theoretical similarities. We also note that the lowest score is for robust modified outcome and GBOOST. In the previous section,

these two methods were the best performing. This suggests those two methods can make different true discoveries.

	GBOOST	Modified outcome	Normalized modified outcome	Shifted modified outcome	Robust modified outcome	Product LASSO
GBOOST	1.000	0.200	0.203	0.202	0.070	0.152
Modified outcome	0.200	1.000	0.411	0.405	0.150	0.283
Normalized modified outcome	0.203	0.411	1.000	0.406	0.153	0.284
Shifted modified outcome	0.202	0.405	0.406	1.000	0.179	0.301
Robust modified outcome	0.070	0.150	0.153	0.179	1.000	0.257
Product LASSO	0.152	0.283	0.284	0.301	0.257	1.000

Table 1: Concordance between methods used to determine SNPs synergistic to rs41475248 in type II diabetes, measured by Kendall's tau.

In any follow-up work, we will only exploit the highly-ranked variants. A weighted tau statistic that assigns a higher weight to the first instances is therefore more relevant. Weighted nonnegative tau statistics better assess the relative level of concordance between different pairs of methods, while the sign in Kendall's tau shows if two methods rather agree or disagree. In Table 2, we list Kendall's tau coefficients with multiplicative hyperbolic weighting. Similarly, we notice that robust modified outcome is least correlated with GBOOST and most correlated with product LASSO.

	GBOOST	Modified outcome	Normalized modified outcome	Shifted modified outcome	Robust modified outcome	Product LASSO
GBOOST	1.000	0.483	0.481	0.517	0.423	0.501
Modified outcome	0.483	1.000	0.851	0.857	0.462	0.586
Normalized modified outcome	0.481	0.851	1.000	0.860	0.467	0.594
Shifted modified outcome	0.517	0.857	0.860	1.000	0.504	0.603
Robust modified outcome	0.423	0.462	0.467	0.504	1.000	0.596
Product LASSO	0.501	0.586	0.594	0.603	0.596	1.000

Table 2: Concordance between methods used to determine SNPs synergistic to rs41475248 in type II diabetes, measured by Kendall's tau with multiplicative weights.

Aside from rank correlation, another option to appraise the results is to measure the association between the top SNPs for each method and the phenotype. Table 3 lists the Cochran-Armitage test p-values for the top 25 SNPs for each method in an increasing order. Though synthetic univariate measures, the Cochran-Armitage statistics give us an indication of the true ranking performance. Robust modified outcome is clearly the method with the lowest p-values. For instance, the top 14 SNPs have a p-value lower than 0.001. That confirms the result of our simulations that robust modified outcome is the best performer for capturing causal SNPs. The p-values associated to product LASSO and

GBOOST are also relatively low, with respectively 5 and 4 p-values lower than 0.001. However, we note the overall difficulty in drawing clear conclusions for all methods. Without multiple testing correction, most of the p-values for each method already exceed classical significance levels *e.g.* 0.05. For 3 out of 6 methods, the p-values of the 25^{th} SNP are greater than 0.90. Nonetheless, the existence of such high p-values further demonstrates the capacity of our methods in discovering novel associations undetected by univariate methods.

GBOOST	Modified outcome	Normalized modified outcome	Shifted modified outcome	Robust modified outcome	Product LASSO
0.0000047	**0.0000000**	**0.0000000**	**0.0000000**	**0.0000000**	**0.0000047**
0.0002632	**0.0000015**	**0.0000015**	**0.0000015**	**0.0000000**	**0.0000075**
0.0002667	**0.0002667**	**0.0002667**	**0.0002667**	**0.0000001**	**0.0000172**
0.0006166	0.0027308	0.0027308	0.0027308	**0.0000012**	**0.0002667**
0.0015069	0.0093734	0.0093734	0.0093734	**0.0000049**	**0.0005286**
0.0028872	0.0633055	0.0633055	0.0633055	**0.0000059**	0.0110392
0.0031533	0.0724198	0.0724198	0.0724198	**0.0000075**	0.0122543
0.0034323	0.0925877	0.0925877	0.0771170	**0.0000172**	0.0152912
0.0081128	0.1126164	0.1043632	0.0925877	**0.0002030**	0.0346055
0.0093734	0.1272777	0.1126164	0.1126164	**0.0002667**	0.0347964
0.0142695	0.2552284	0.1567974	0.1272777	**0.0003047**	0.0396448
0.0633055	0.2926915	0.2971396	0.1639805	**0.0004643**	0.0396932
0.0771170	0.3436741	0.3529366	0.2971396	**0.0005286**	0.0527104
0.1616393	0.3529366	0.5012038	0.3529366	**0.0005841**	0.0633055
0.2089538	0.5871432	0.5506690	0.5012038	0.0015214	0.0763114
0.2114803	0.5985624	0.5985624	0.5707955	0.0016353	0.1126164
0.2256368	0.6016953	0.7183847	0.5985624	0.0025709	0.1185275
0.2586186	0.6361937	0.7199328	0.7000506	0.0064196	0.1796624
0.2654530	0.7183847	0.7342897	0.7183847	0.0080405	0.2552284
0.4105146	0.7342897	0.7656055	0.7342897	0.0110392	0.3308890
0.4323674	0.7979653	0.7706524	0.7979653	0.0122543	0.3867409
0.4376669	0.8683271	0.7979653	0.7993838	0.0124442	0.5045073
0.4796214	0.8820292	0.7993838	0.8683271	0.0136452	0.5985624
0.5871432	0.9188037	0.8820292	0.8821872	0.0346055	0.6238335
0.9479547	0.9903334	0.8821872	0.9188037	0.0396932	0.8821872

Table 3: Cochran-Armitage test p-values for the top 25 SNPs for each method

4 Discussion

We presented a new family of methods for epistasis detection. They revolve around detecting new interactions with specific targets/genes. Given our partial understanding of common diseases, such refocused models could be more useful in the understanding of the underlying biology. Hundreds of genes have already been associated with several common diseases via univariate GWAS. For type II diabetes, we mention the genes TCF7L2 and ABCC8. The latter affects insulin regulation, while the former impacts both insulin secretion and glucose production. The next step is to

build upon these findings to detect potential synergies between these genes and the rest of the genome. Beyond a better understanding of disease mechanisms through new biomarker discovery, we see the development of combination drug therapies as a potential application of our work.

Among the methods we propose, robust modified outcome seems the most suited in practice to GWAS applications. The AUCs are overall the highest in addition to the early retrieval performance. More importantly, robust modified outcome outperforms GBOOST. From a dimensionality point of view, the closest simulations to real GWAS are for sample sizes $n = 500$. Across the four scenarios, robust modified outcome not only outperforms the current state-of-the art for epistasis detection GBOOST, but also the other methods based on regression models. However, the low PR AUCs show that there is still room for improvement. The highest observed PR AUC is 0.17. In the PR space, we also note that several of our methods clearly outperform GBOOST for all scenarios and all sample sizes. Interestingly, the GBOOST ROC curves behave similarly to other methods. Such differences between ROC and PR curves are common for highly-skewed datasets where PR curves are more informative[53]. The main point of our methods is to focus on the synergies with a particular target while discarding other effects. The consistent ROC and PR AUCs across the four different scenarios show that they are rather successful at that. Their performance is not strongly impacted by the presence of additional marginal and/or epistatic effects.

The case study that we carried for type II diabetes demonstrates the scalability of all methods to real GWAS. One way to improve runtime is to adjust the number of subsamples used for stability selection; however this may come at the expense of performance. The development of new and faster LASSO solvers[54,55] for large scale problems will further help improve the adoption of our methods by end-users.

According to two rank correlation measures (Kendall's tau and weighted Kendall's tau), we see that all methods tend to agree, though partially. In simulations, synthetic performance measures like ROC and PR AUCs were relatively close. On the other hand, the rank correlations do not show complete agreement (values far from 1). For instance, GBOOST least agrees with robust modified outcome. However, the two methods are the best performing in our simulations. We conclude that a consensus method combining GBOOST and robust modified outcome could improve the recovery of interacting SNPs. Theoretically, the ranking differences between the methods motivate the question of the guarantees for support recovery in terms of effect sizes and dependence structure among covariates. Common variants with low effect sizes is a major hypothesis for missing heritability. A recent paper from Boyle et al.[56] even advances the hypothesis of an "omnigenic" model. It proposes that most heritability lies outside of core pathways; principally within regulatory pathways. That means that a large number of variants influence the phenotype. However, that brings up the question of causality: how to define a causal SNP when all variants are related to phenotype?

The simulations prove that a number of the highly-ranked SNPs are false positives. That is accentuated by the imbalanced nature of our problem: a handful of causal SNPs for thousands of referenced SNPs. Hopefully, the continual decrease in genotyping costs will result in a dramatic increase in sample sizes and, in consequence, statistical

power. For instance, the UK Biobank[57] comprises full genome-wide data for five hundred thousand individuals. We also point out that our methods do naturally extend to higher-order interactions. The main idea is combining two SNPs into a single target through a binary function such as the product of the two SNPs. We expect results to depend on both the combination rule and our encoding choice for each SNP. Moreover, a loss of information occurs with such simplifications. We leave a study of those extensions to future work.

The main contribution of our work is extending the causal inference framework to epistasis by developing propensity-like scores for genomic data. The superior performance of robust modified outcome is partially owed to its robustness against propensity scores misspecification. An area of improvement is the propensity score estimation which can benefit a large number of methods. An interesting proposal from Wager et al.[58] completely forgoes propensity scores for the estimate of average treatment effects. All of the presented methods were originally developed for clinical trials where the analog to the target SNP is the treatment assignment and to the genotype are the clinical covariates. Given the rich literature in that field, this opens the door to a much broader panel of methods. In particular, future directions of our work include conditioning for multiple covariates (whether clinical covariates, variables encoding population structure or other genetic variants) to account for, among other things, higher-order interactions and population stratification.

Acknowledgements

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from wwww.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

References

[1] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

[2] McCarthy, M. I. and Hirschhorn, J. N. (2008). Genome-wide association studies: potential next steps on a genetic journey. Human Molecular Genetics *17*, R156–R165.

[3] Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America *109*, 1193–8.

[4] Zeggini, E., , Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature Genetics *40*, 638–645.

[5] Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., et al. (2008). Many sequence variants affecting diversity of adult human height. Nature Genetics *40*, 609–615.

[6] Combarros, O., Cortina-Borja, M., Smith, A. D., and Lehmann, D. J. (2009). Epistasis in sporadic alzheimer's disease. Neurobiology of Aging *30*, 1333–1349.

[7] Fisher, R. A. (1919). XV.—the correlation between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburgh *52*, 399–433.

[8] Cordell, H. J. (2009). Detecting genegene interactions that underlie human diseases. Nature Reviews Genetics *10*, 392–404.

[9] Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. Frontiers in Genetics *6*.

[10] Moore, J. H. and Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays *27*, 637–646.

[11] Emily, M. (2016). AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies. Statistical Applications in Genetics and Molecular Biology *15*.

[12] Stanislas, V., Dalmasso, C., and Ambroise, C. (2017). Eigen-epistasis for detecting gene-gene interactions. BMC Bioinformatics *18*.

[13] Cabin, R. J. and Mitchell, R. J. (2000). To bonferroni or not to bonferroni: when and how are the questions. Bulletin of the Ecological Society of America *81*, 246–248.

[14] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289–300.

[15] Nakagawa, S. (2004). A farewell to bonferroni: the problems of low statistical power and publication bias. Behavioral Ecology *15*, 1044–1045.

[16] Chatelain, C., Durand, G., Thuillier, V., and Augé, F. (2018). Performance of epistasis detection methods in semi-simulated GWAS. BMC Bioinformatics *19*.

[17] Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., and Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. American Journal of Human Genetics *87*, 325–340.

[18] Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). GBOOST: a GPU-based tool for detecting genegene interactions in genomewide case control studies. Bioinformatics *27*, 1309–1310.

[19] Thanei, G.-A., Meinshausen, N., and Shah, R. D. (2018). The xyz algorithm for fast interaction search in high-dimensional data. Journal of Machine Learning Research *19*, 1–42.

[20] Winkelmann, B. R., März, W., Boehm, B. O., Zotz, R., Hager, J., Hellstern, P., and Senges, J. (2001). Rationale and design of the LURIC study - a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. Pharmacogenomics *2*, S1–S73.

[21] Tibshirani, R., Johnstone, I., Hastie, T., and Efron, B. (2004). Least angle regression. The Annals of Statistics *32*, 407–499.

[22] Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. The Annals of Statistics *41*, 1111–1141.

[23] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology *66*, 688–701.

[24] Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics *78*, 629–44.

[25] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *72*, 417–473.

[26] Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. Journal of the American Statistical Association *107*, 1106–1118.

[27] Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. Journal of the American Statistical Association *109*, 1517–1532.

[28] Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine *23*, 2937–2960.

[29] Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine *34*, 3661–3679.

[30] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics *81*, 559–575.

[31] Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P., and Becker, T. (2009). Intersnp: genome-wide interaction analysis guided by a priori information. Bioinformatics *25*, 3275–3281.

[32] Gatto, N. M. (2004). Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. International Journal of Epidemiology *33*, 1014–1024.

[33] Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statistics in Medicine *13*, 153–162.

[34] Yang, Q., Khoury, M. J., Sun, F., and Flanders, W. D. (1999). Case-only design to measure gene-gene interaction. Epidemiology (Cambridge, Mass.) *10*, 167–70.

[35] Sun, S., Greenwood, C. M., and Neal, R. M. (2007). Haplotype inference using a bayesian hidden markov model. Genetic Epidemiology *31*, 937–948.

[36] Rastas, P., Koivisto, M., Mannila, H., and Ukkonen, E. (2005). A hidden markov technique for haplotype reconstruction. In Lecture Notes in Computer Science In Lecture Notes in Computer Science. (Springer Berlin Heidelberg).

[37] Kimmel, G. and Shamir, R. (2005). A block-free hidden markov model for genotypes and its application to disease association. Journal of Computational Biology *12*, 1243–1260.

[38] Gabriel, S. B. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225–2229.

[39] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics *43*, 2055–2085.

[40] Sesia, M., Sabatti, C., and Candès, E. J. (2018). Gene hunting with hidden markov model knockoffs. Biometrika.

[41] Cole, S. R. and Hernan, M. A. (2008). Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology *168*, 656–664.

[42] Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. PLoS ONE *6*, e18174.

[43] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE *77*, 257–286.

[44] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *67*, 301–320.

[45] Haury, A. C., Mordelet, F., Vera-Licona, P., and Vert, J. P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. BMC Systems Biology *6*.

[46] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software *33*.

[47] Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development *19*, 212–219.

[48] Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.

[49] Auton, A. e. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

[50] Burton, P. R. et al. (2007). Genome-wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature *447*, 661–678.

[51] Saito, T. and Rehmsmeier, M. (2016). Precrec: fast and accurate precision–recall and ROC curve calculations in r. Bioinformatics *33*, 145–147.

[52] Zeng, Y. and Breheny, P. (2017). The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in r. ArXiv e-prints.

[53] Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML '06 pp. 233–240.

[54] Le Morvan, M. and Vert, J. (2018). WHInter: A working set algorithm for high-dimensional sparse second order interaction models. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 pp. 3632–3641.

[55] Massias, M., Gramfort, A., and Salmon, J. (2018). Celer: a Fast Solver for the Lasso with Dual Extrapolation. In ICML 2018 - 35th International Conference on Machine Learning volume 80 of *PMLR* pp. 3321–3330.

[56] Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. Cell *169*, 1177–1186.

[57] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2017). Genome-wide genetic data on 500,000 uk biobank participants. bioRxiv.

[58] Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B (Statistical Methodology).

[59] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) *39*, 1–38.

A Genotypic hidden Markov model

In this Appendix, we explicit the transition and emission probabilities for the genotypic hidden Markov model. For that purpose, we start by considering a pair of ordered haplotypes $H^a = \left(H_1^a, \cdots, H_p^a\right) \in \{0,1\}^p$ and $H^b = \left(H_1^b, \cdots, H_p^b\right) \in \{0,1\}^p$. We recall that the two haplotypes correspond to the same positions. The hidden variables $Z^a = \left(Z_1^a, \cdots, Z_p^a,\right)$ and $Z^b = \left(Z_1^b, \cdots, Z_p^b,\right)$ represent cluster memberships. They take discrete values in $\{1, \cdots, K\}^p$. Scheet and Stephens[24] define the clusters as a "(common) combination of alleles at tightly linked SNPs". The underlying hidden Markov models for the two alleles have identical forms. We then focus on the first allele a. We follow the notations of[40].

The marginal distribution of the first hidden state can be written as:

$$q_1^{hap}(k) = \alpha_{1,k}, \quad k \in \{1, \cdots, K\}.$$

For $j \in \{2, \cdots, p\}$, the transition matrix Q_j^{hap} is given by:

$$Q_j^{hap}(k'|k) = P(H_j = k'|H_{j-1} = k') = \begin{cases} e^{-r_j} + (1 - e^{-r_j})\, \alpha_{j,k'}, & k' = k \\ (1 - e^{-r_j})\, \alpha_{j,k'}, & k' \neq k \end{cases}.$$

The parameter $r = (r_2, \cdots, r_p)$ can be assimilated to the recombination rate between loci $j-1$ and j, although Scheet and Stephens[24] point out the general mismatch between the observed recombination rates and the estimate of r. The parameter $\alpha = (\alpha_{j,k})_{(j,k) \in \{1,\cdots p\} \times \{1,\cdots,K\}}$ is the relative frequency of the cluster k in locus j.

Conditionally on the latent state $Z_j^{hap} = z_j$, the allele H_j is a Bernoulli random variable, $H_j|Z_j \sim \mathcal{B}(\theta_{j,z_j})$. θ_{j,z_j} is the frequency of allele 1 in cluster z_j at the position j:

$$f_j^{hap} = (h_j; z_j, \theta) = \begin{cases} 1 - \theta_{j,z_j}, & h_j = 0 \\ \theta_{j,z_j}, & h_j = 1 \end{cases}.$$

Under the Hardy-Weinberg equilibrium (HWE), a third hidden Markov model for the unphased genotype can be derived by combining the HMMs of the two alleles a and b. The emission states $X = (X_1, \cdots, X_p) \in \{0,1,2\}^p$ are given by the sum of the emission states, $H^a + H^b = \left(H_1^a + H_1^b, \cdots, H_p^a + H_p^b\right)$. Because of the phase indetermination, the latent states are unordered pairs of haplotype latent states, $Z = (\{Z_1^a, Z_1^b\}, \cdots, \{Z_p^a, Z_p^b\})$. Thus, the dimensionality of the latent variable space is $K(K+1)/2$. The different probabilities of the genotype model are computed by considering the two cases: $Z_j^a = Z_j^b$ and $Z_j^a \neq Z_j^b$.

The initial latent state distribution is given by:

$$q_1^{gen}(\{k^a, k^b\}) = \begin{cases} (\alpha_{1,k^a})^2, & k^a = k^b \\ 2\alpha_{1,k^a}\alpha_{1,k^b} & k^a \neq k^b \end{cases},$$

In a similar fashion, the transition probabilities:

$$Q_j^{gen}(\{\underline{k}^a, \underline{k}^b\}|\{k^a, k^b\}) = \begin{cases} Q_j^{hap}(\underline{k}^a|k^a)Q_j^{hap}(\underline{k}^b|k^b) + Q_j^{hap}(\underline{k}^b|k^a)Q_j^{hap}(\underline{k}^a|k^b), & \underline{k}^a \neq \underline{k}^b \\ Q_j^{hap}(\underline{k}^a|k^a)Q_j^{hap}(\underline{k}^b|k^b), & \text{otherwise} \end{cases},$$

and, the emission probabilities are

$$f_j(x_j; \{k^a, k^b\}, \theta) = \begin{cases} (1-\theta_{j,k^a})(1-\theta_{j,k^b}), & x_j = 0 \\ \theta_{j,k^a}(1-\theta_{j,k^b}) + (1-\theta_{j,k^a})\theta_{j,k^b}, & x_j = 1 \\ \theta_{j,k^a}\theta_{j,k^b}, & x_j = 2 \end{cases}.$$

For the estimate of the parameters $\nu = (\alpha, r, \theta)$, we use the imputation software fastPHASE[24] which fits the hidden Markov model using an expectation-maximization (EM) algorithm[59]. Its computational complexity is $\mathcal{O}\left(npK^2\right)$. The complexity scales linearly for both p and n, rendering fastPHASE well-suited for real case-control datasets where the number of SNPs is typically in the hundreds of thousands and the number of samples in the thousands. In practice, as a trade-off between a rich representation of the clusters and the ensuing quadratic complexity, we chose $K = 12$.

B Simulation results

B.1 First scenario: synergistic only effects

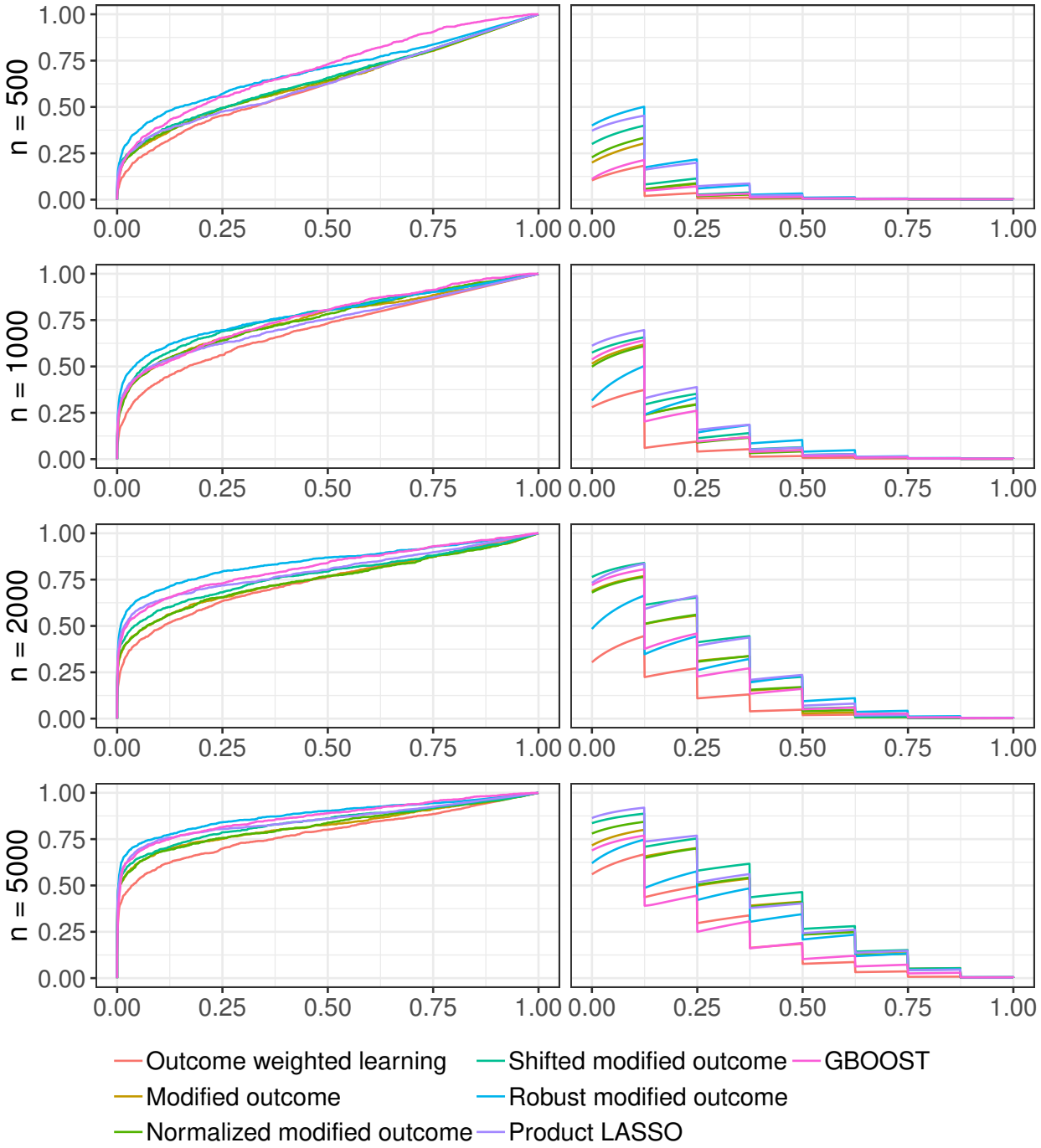

Figure 2: Average ROC (left column) and PR (right column) curves for the first scenario

Table 4: Average ROC and PR AUCs for the first scenario

Method	PR	ROC
n =500		
GBOOST	0.0362	0.7075
Modified outcome	0.0468	0.6747
Robust modified outcome	0.0973	0.7414
Normalized modified outcome	0.0512	0.6754
Shifted modified outcome	0.0644	0.6794
Outcome weighted learning	0.0254	0.6282
Product LASSO	0.0895	0.6514
n =1000		
GBOOST	0.1270	0.7688
Modified outcome	0.1284	0.7131
Robust modified outcome	0.1302	0.7434
Normalized modified outcome	0.1255	0.7120
Shifted modified outcome	0.1470	0.7224
Outcome weighted learning	0.0613	0.6764
Product LASSO	0.1619	0.7032
n =2000		
GBOOST	0.2103	0.8169
Modified outcome	0.2252	0.7512
Robust modified outcome	0.2070	0.8449
Normalized modified outcome	0.2266	0.7501
Shifted modified outcome	0.2704	0.7753
Outcome weighted learning	0.1045	0.7394
Product LASSO	0.2711	0.7989
n =5000		
GBOOST	0.2276	0.8697
Modified outcome	0.3512	0.8218
Robust modified outcome	0.3011	0.8818
Normalized modified outcome	0.3548	0.8248
Shifted modified outcome	0.3907	0.8423
Outcome weighted learning	0.2139	0.7847
Product LASSO	0.3779	0.8546

B.2 Second scenario: partial overlap between synergistic and marginal effects

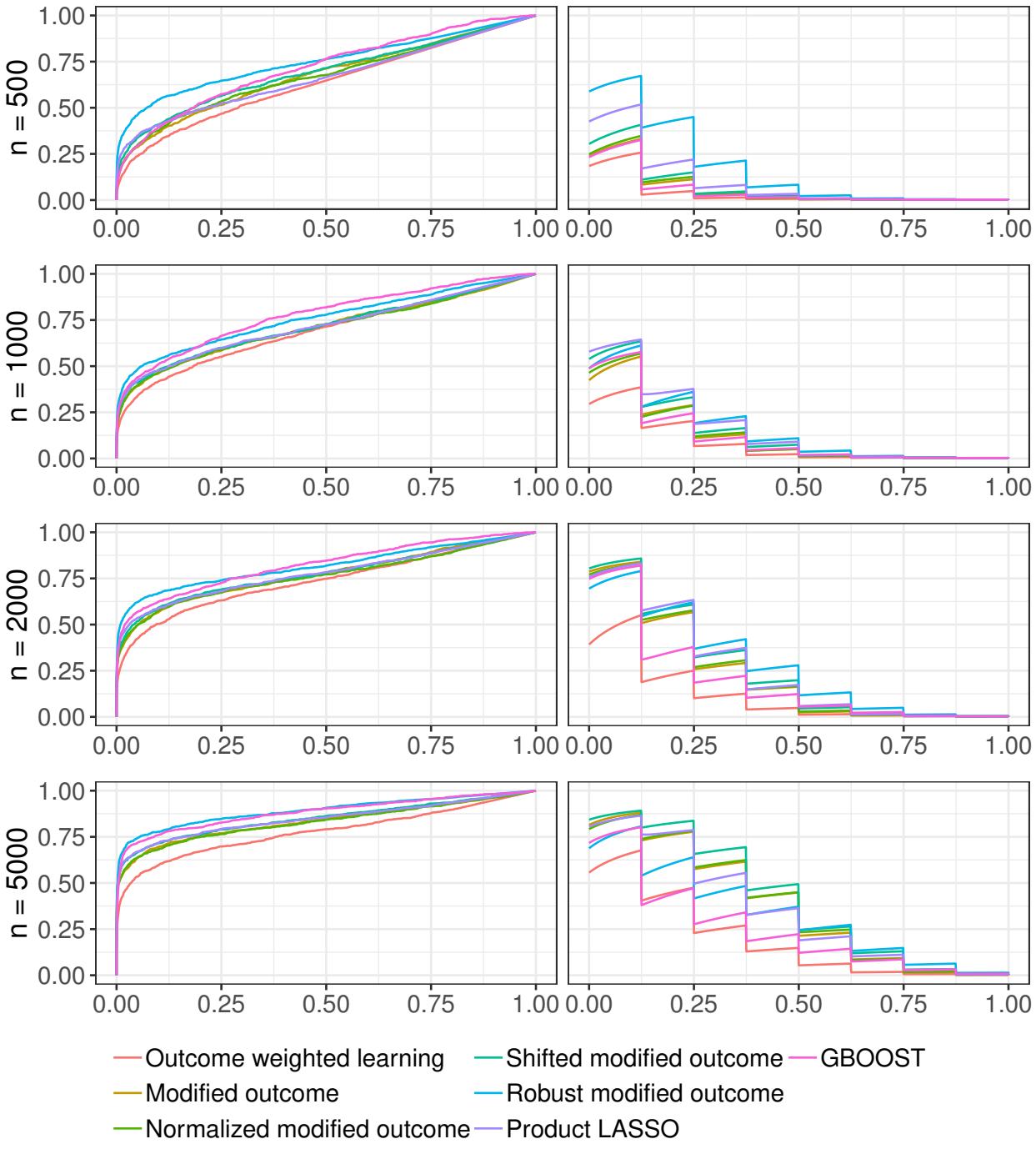

Figure 3: Average ROC (left column) and PR (right column) curves for the second scenario

Table 5: Average ROC and PR AUCs for the second scenario

Method	PR	ROC
n =500		
GBOOST	0.0516	0.7186
Modified outcome	0.0563	0.6750
Robust modified outcome	0.1716	0.7502
Normalized modified outcome	0.0590	0.6713
Shifted modified outcome	0.0712	0.6918
Outcome weighted learning	0.0367	0.6345
Product LASSO	0.0994	0.6659
n =1000		
GBOOST	0.1190	0.7773
Modified outcome	0.1195	0.7092
Robust modified outcome	0.1574	0.7601
Normalized modified outcome	0.1233	0.7080
Shifted modified outcome	0.1443	0.7160
Outcome weighted learning	0.0805	0.6923
Product LASSO	0.1609	0.7170
n =2000		
GBOOST	0.1933	0.8226
Modified outcome	0.2294	0.7708
Robust modified outcome	0.2732	0.8183
Normalized modified outcome	0.2321	0.7623
Shifted modified outcome	0.2532	0.7753
Outcome weighted learning	0.1114	0.7360
Product LASSO	0.2507	0.7762
n =5000		
GBOOST	02454	0.8821
Modified outcome	0.3718	0.8344
Robust modified outcome	0.3286	0.8916
Normalized modified outcome	0.3739	0.8309
Shifted modified outcome	0.4079	0.8487
Outcome weighted learning	0.1930	0.7769
Product LASSO	0.3537	0.8467

B.3 Third scenario: partial overlap between synergistic and quadratic effects

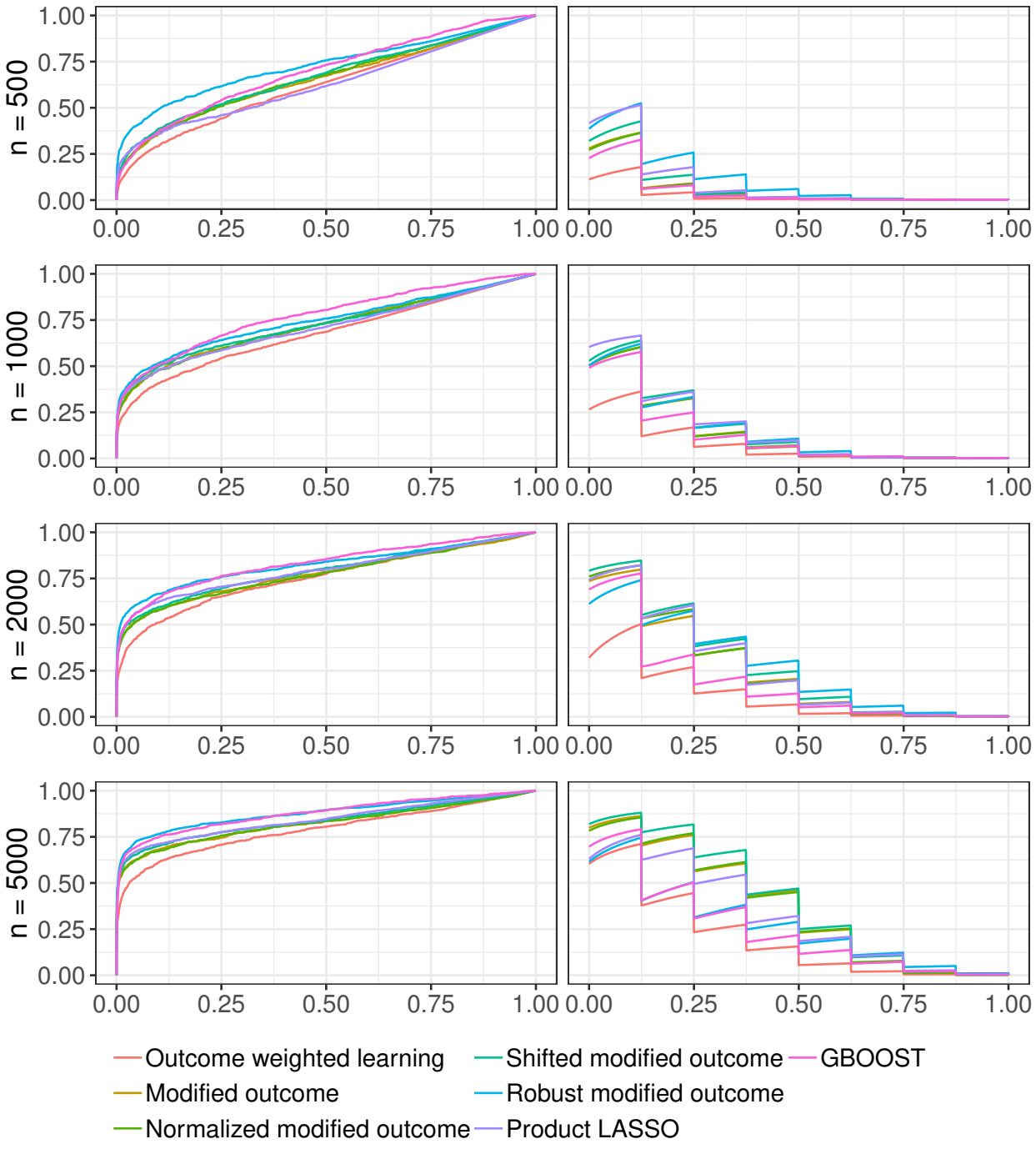

Figure 4: Average ROC (left column) and PR (right column) curves for the third scenario

Table 6: Average ROC and PR AUCs for the third scenario

Method	PR	ROC
n =500		
GBOOST	0.050	0.6970
Modified outcome	0.0570	0.6559
Robust modified outcome	0.1148	0.7296
Normalized modified outcome	0.0569	0.6627
Shifted modified outcome	0.0714	0.6703
Outcome weighted learning	0.0260	0.6233
Product LASSO	0.0889	0.6282
n =1000		
GBOOST	0.1228	0.7746
Modified outcome	0.1362	0.7181
Robust modified outcome	0.1513	0.7444
Normalized modified outcome	0.1373	0.7175
Shifted modified outcome	0.1546	0.7226
Outcome weighted learning	0.0728	0.6778
Product LASSO	0.1620	0.7100
n =2000		
GBOOST	0.1814	0.8307
Modified outcome	0.2430	0.7733
Robust modified outcome	0.2697	0.8235
Normalized modified outcome	0.2496	0.7724
Shifted modified outcome	0.2737	0.7886
Outcome weighted learning	0.1129	0.7535
Product LASSO	0.2543	0.7921
n =5000		
GBOOST	0.2467	0.8767
Modified outcome	0.3663	0.8241
Robust modified outcome	0.2660	0.8790
Normalized modified outcome	0.3669	0.8236
Shifted modified outcome	0.3944	0.8376
Outcome weighted learning	0.1965	0.7893
Product LASSO	0.3158	0.8439

B.4 Fourth scenario: partial overlap between synergistic and quadratic/marginal effects

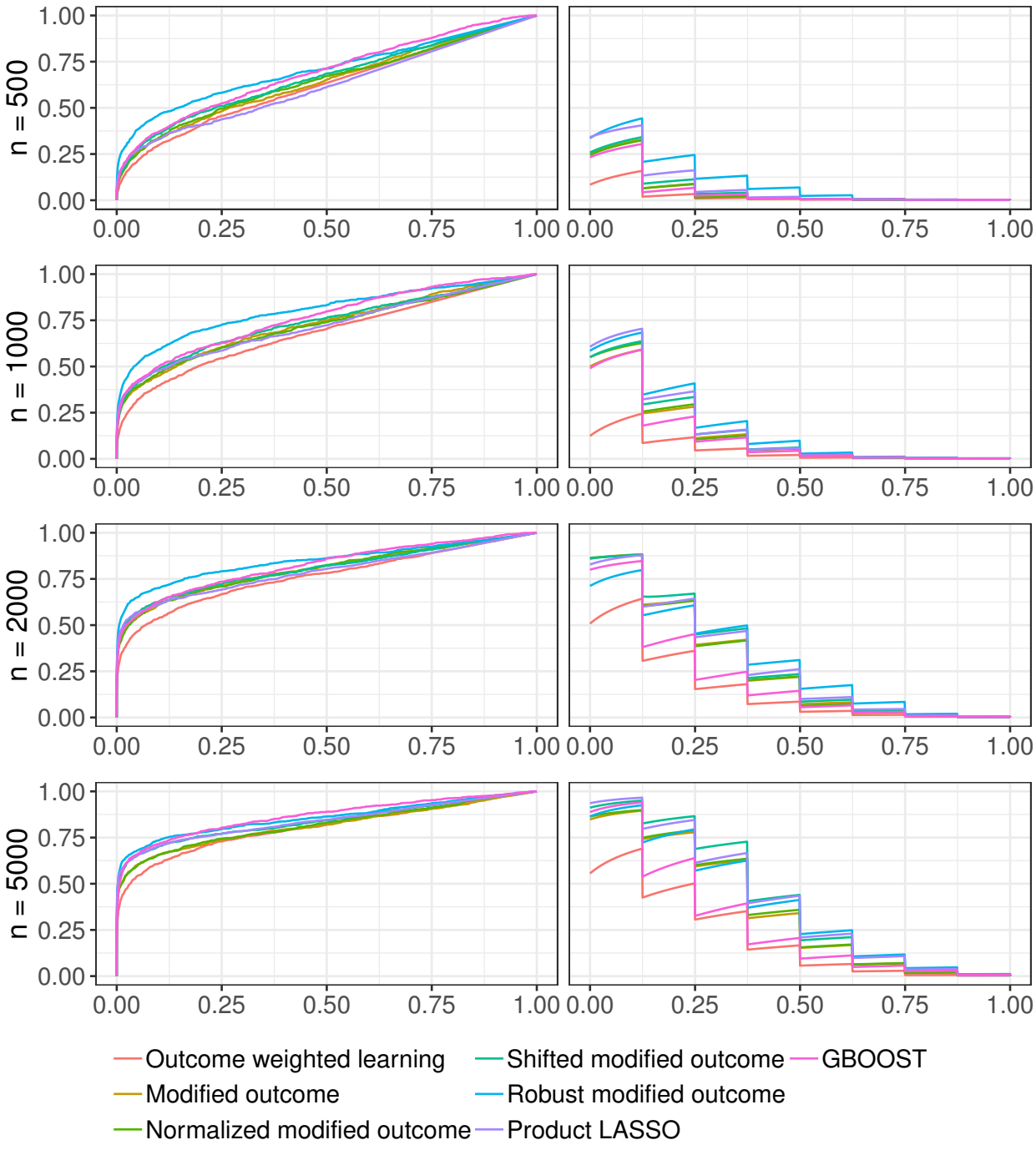

Figure 5: Average ROC (left column) and PR (right column) curves for the fourth scenario

Table 7: Average ROC and PR AUCs for the fourth scenario

Method	PR	ROC
n =500		
GBOOST	0.0479	0.6900
Modified outcome	0.0521	0.6427
Robust modified outcome	0.1066	0.7065
Normalized modified outcome	0.0513	0.6460
Shifted modified outcome	0.0591	0.6623
Outcome weighted learning	0.0227	0.6218
Product LASSO	0.0762	0.6174
n =1000		
GBOOST	0.1163	0.7647
Modified outcome	0.1283	0.7288
Robust modified outcome	0.1687	0.8049
Normalized modified outcome	0.1338	0.7200
Shifted modified outcome	0.1438	0.7388
Outcome weighted learning	0.0479	0.6838
Product LASSO	0.1554	0.7206
n =2000		
GBOOST	0.2129	0.8237
Modified outcome	0.2794	0.8007
Robust modified outcome	0.2986	0.8478
Normalized modified outcome	0.2763	0.8032
Shifted modified outcome	0.2960	0.8050
Outcome weighted learning	0.1530	0.7641
Product LASSO	0.2927	0.7899
n =5000		
GBOOST	0.2823	0.8656
Modified outcome	0.3541	0.8127
Robust modified outcome	0.3823	0.8568
Normalized modified outcome	0.3597	0.8175
Shifted modified outcome	0.4091	0.8388
Outcome weighted learning	0.2106	0.8031
Product LASSO	0.4000	0.8399