Nearest-neighbor Projected-Distance Regression (NPDR) detects network interactions and controls for confounding and multiple testing

Trang T. Le[1], Bryan A. Dawkins[2] and Brett A. McKinney[2,3*]

[1]Department of Biostatistics, Epidemiology and Informatics, University of
Pennsylvania, Philadelphia, PA 19104
[2]Department of Mathematics, University of Tulsa, Tulsa, OK 74104
[3]Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104

March 7, 2019

Abstract

Efficient machine learning methods are needed to detect complex interaction network effects in complicated modeling scenarios in high dimensional data, such as GWAS or gene expression for case-control or continuous outcomes. Many machine learning feature selection methods have limited ability to address the issues of controlling the false discovery rate and adjusting for covariates. To address these challenges, we develop a new feature selection technique called Nearest-neighbor Projected-Distance Regression (NPDR) that uses the generalized linear model (GLM) to perform regression between nearest-neighbor pair distances projected onto predictor dimensions. Motivated by the nearest-neighbor mechanism in Relief-based algorithms, NPDR captures the underlying interaction structure of the data, handles both dichotomous and continuous outcomes and various combinations of predictor data types, statistically corrects for covariates and allows for regularization. Using realistic simulations with main effects and network interactions, we show that NPDR outperforms standard Relief-based methods and random forest at detecting functional variables while also enabling covariate adjustment and multiple testing correction. Using RNA-Seq data from a study of major depressive disorder, we show that NPDR with covariate adjustment effectively removes spurious associations due to confounding. We apply NPDR to a separate RNA-Seq study with a continuous outcome of sleep quality and identify genes important to the phenotype.

1 Introduction

Epistasis is a measure of the effect of a genetic variant on a phenotype beyond what would be expected by their independent effects. There is evidence that these non-independent effects are pervasive [1] and that higher-order interactions also play an important role in genetics [2]. A similar interaction effect can be observed in differential co-expression, where the phenotypic effect of one gene is modified depending on the expression of another gene [3,4]. The embedding of these interactions in a regulatory network may lead to, not only pairwise interactions, but also higher-order epistasis network effects. Attempting to model these higher order interactions explicitly would be computationally and statistically intractable [5]. Thus, computationally scalable feature

selection methods are needed to capture these higher-order effects in high-dimensional data such as genome-wide association [6] and gene expression studies [7].

Relief-based algorithms are efficient nearest-neighbor feature selection methods that are able to detect epistasis or statistical interaction effects in high-dimensional data without resorting to pairwise modeling of attributes [8–11]. We recently introduced the STatistical Inference Relief (STIR) formalism [7] to address Relief-based methods' lack of a statistical distribution for hypothesis testing and the challenge of assessing the false positive rate of Relief-based scores. STIR extended Relief-based methods to compute statistical significance of attributes in binary outcome data (e.g., case-control) by reformulating the Relief weight [12] as a pseudo t-test for the difference of means between projected-distances onto attributes between neighbors with the same and opposite phenotype (hits and misses). We showed that STIR is an effective approach with high power and low false-positive rates for data with main and interaction effects. STIR is applicable to any predictor data type (continuous/expression or nominal/variants); however, being based on a t-test, it does not apply to data with a continuous outcome (e.g., quantitative trait) and does not correct for covariates. Prior to the current study, no Relief-based method included covariate correction.

In the current study, we introduce a new Nearest-neighbor Projected-Distance Regression (NPDR) approach that extends the STIR formalism to regression of nearest neighbors with the generalized linear model (GLM). For each attribute, the NPDR model fits a GLM of projected-distance differences (onto the attribute) between all pairs of nearest-instance neighbors. The model is linear for regression problems and logistic for dichotomous outcome problems, and either model may include terms for covariates. The importance of an attribute is given by its standardized regression coefficient and the statistical significance by its P value.

The flexible GLM formalism of NPDR opens up Relief-based methods to statistical inference for a broad class of problems. It leads to an improved attribute importance estimator for continuous outcome problems compared to other regression Relief methods, adjusts for covariates (Eq. 11), and allows for hypothesis testing, while detecting main effects and interactions. The model gives the appearance of being univariate but implicitly accounts for interactions with all other attributes via the neighborhood calculation in the space of all attributes (omnigenic). NPDR is applicable to any predictor data types such as SNPs in GWAS or expression in RNA-Seq analyses. It allows for hypothesis testing and multiple testing adjustment. Further, we implement a penalized version of NPDR with the elasticnet penalty.

Covariate adjustment is often neglected in machine learning, yet many biological and clinical studies involve potentially confounding covariates such as sex, bmi, and age [13] or population stratification [14]. Some proposed methods for correcting machine learning algorithms include restricted permutation [15], inverse probability weighting of training samples [16] and penalized support vector machines [17]. Our NPDR framework leads to a natural way to control for covariates by including additional terms for the between-neighbor projected differences for each covariate within the GLM. We demonstrate the effectiveness of NPDR to correct for confounding in an RNA-Seq study of major depressive disorder in which there is a strong signal in the expression data due to the sex of study participants.

The paper is organized as follows. In the Methods section, we develop the new formalism of NPDR to reformulate Relief-based scores as coefficients in a distanced-based GLM. We use the projected-distance regression formalism to implement a penalized version of NPDR. In the Results, we use realistic simulations with main effects and network interactions to demonstrate improved feature selection performance over standard Relief-based methods and random forest. We apply NPDR with covariate correction to two real RNA-Seq datasets from studies of major depressive disorder

(MDD), including dichotomous outcome (MDD diagnosis status) and continuous 63
outcome (Pittsburgh Sleep Quality Index, PSQI). We show that NPDR removes 64
spurious associations due to confounding by sex and reveals genes important to the 65
phenotypes. 66

2 Materials and Methods 67

In this section, we develop the mathematical formalism needed to describe the projected 68
distance regression in NPDR. We then construct the NPDR GLM models for common 69
analysis situations, including continuous and dichotomous outcomes and adjustment for 70
covariates. We also describe the simulation approach and real data sets for method 71
validation. 72

2.1 Distance metrics and nearest neighbors 73

Because NPDR and other Relief-based feature selection methods are based on distances 74
between instances, we first describe the algorithms and notation for identifying nearest 75
neighbors in the space all attributes. We use the term attribute to refer to predictor 76
variables, which may be continuous (e.g., expression) or categorical (e.g., variants). We 77
use the term instance to refer to samples or subjects in a dataset. 78

2.1.1 Distances and projections onto attributes 79

The distance between instances i and j in the data set $X^{m \times p}$ of m instances and p 80
attributes is calculated in the space of all attributes ($a \in A$, $|A| = p$) using a metric 81
such as 82

$$D_{ij}^{(q)} = \left(\sum_{a \in A} |\mathrm{d}_{ij}(a)|^q \right)^{1/q}, \qquad (1)$$

which is typically Manhattan ($q = 1$) but may also be Euclidean ($q = 2$). The quantity 83
$\mathrm{d}_{ij}(a)$, known as a "diff" in Relief literature, is the projection of the distance between 84
instances i and j onto the attribute a dimension. The function $\mathrm{d}_{ij}(a)$ supports any type 85
of attributes (e.g., numeric continuous versus categorical). For example, the projected 86
difference between two instances i and j for a continuous numeric ($\mathrm{d}^{\mathrm{num}}$) attribute a 87
may be 88

$$\begin{aligned} \mathrm{d}_{ij}^{\mathrm{num}}(a) &= \mathrm{diff}(a, (i, j)) \\ &= |\hat{X}_{ia} - \hat{X}_{ja}|, \end{aligned} \qquad (2)$$

where $\hat{X}$ represents the standardized data matrix X. We use a simplified $\mathrm{d}_{ij}(a)$ 89
notation in place of the $\mathrm{diff}(a, (i, j))$ notation that is customary in Relief-based methods. 90
We omit the division by $\max(a) - \min(a)$ used by Relief to constrain scores to the 91
interval from -1 to 1. As we show in subsequent sections, NPDR scores are 92
standardized regression coefficients with corresponding P values, so any scaling 93
operation at this stage is unnecessary for comparing attribute scores. The numeric 94
$\mathrm{d}_{ij}^{\mathrm{num}}(a)$ projection is simply the absolute difference between row elements i and j of the 95
data matrix $X^{m \times p}$ for the attribute column a. 96
    This numeric projection function (Eq. 2) is appropriate for gene expression and other 97
quantitative predictors and outcomes. For genome-wide association study (GWAS) data, 98
where attributes are categorical, one simply modifies the type in the projection 99
function [6], but the projected-distance regression methods will be otherwise unchanged. 100
The $\mathrm{d}_{ij}(a)$ quantity is typically part of the metric to define the neighborhood, but it is 101
also essential for computing the importance coefficients (Sec. 2.2.1). The regression 102

3

models below (Eqs. 8, 11, 15) will be fit for all nearest neighbors i and j in the defined neighborhood (discussed next in Eq. (3)).

2.1.2 Nearest-neighbor ordered pairs

In the original Relief-F approach for binary outcome data, two neighborhood sets are calculated: one for hits and one for misses. For NPDR, only one neighborhood is needed, regardless of whether the problem is classification or regression and regardless of whether a fixed-k or adaptive radius neighborhood method is used [12,18,19]. The NPDR neighbors are chosen blind to the outcome variable and then pairs of instances are assigned to hit or miss groups (instances in the same class or different class) for binary outcome data and assigned numeric differences for quantitative outcome data. This blinded selection leads to less overfitting of the neighborhood boundaries and less bias for imbalanced data.

We define the NPDR neighborhood set $\mathcal{N}$ of ordered pair indices as follows. Instance i is a point in p dimensions, and we designate the topological neighborhood of i as N_i. This neighborhood is a set of other instances trained on the data $X^{m \times p}$ and depends on the type of Relief neighborhood method (e.g., fixed-k or adaptive radius) and the type of metric (e.g., Manhattan or Euclidean). If instance j is in the neighborhood of i ($j \in N_i$), then the ordered pair $(i, j) \in \mathcal{N}$ for the projected-distance regression analysis. The ordered pairs constituting the neighborhood can then be represented as nested sets:

$$\mathcal{N} = \{\{(i, j)\}_{i=1}^{m}\}_{\{j \neq i : j \in N_i\}}. \tag{3}$$

The cardinality of the set $\{j \neq i : j \in N_i\}$ is k_i, the number of nearest neighbors for subject i.

2.1.3 Adaptive-radius and fixed-k Neighborhoods

The NPDR algorithm applies to any Relief neighborhood algorithm. In the applications in the current study, we use the multiSURF [19] adaptive radius neighborhood, which uses a different radius for each instance, and we use a fixed-k neighborhood that well-approximates mulitSURF, which is derived in Ref. [20]. The multiSURF radius for an instance is the mean of its distances to all other instances subtracted by $\alpha = 1/2$ of the standard deviation of this mean. An instance j is in the adaptive α-radius neighborhood of i ($j \in N_i^{\alpha}$) under the condition

$$D_{ij} \leq R_i^{\alpha} \implies j \in N_i^{\alpha}, \tag{4}$$

where the threshold radius for instance i is

$$R_i^{\alpha} = \bar{D}_i - \alpha \, \sigma_{\bar{D}_i} \tag{5}$$

and

$$\bar{D}_i = \frac{1}{m-1} \sum_{j \neq i} D_{ij}^{(\cdot)} \tag{6}$$

is the average of instance i's pairwise distances (using Eq. 1) with standard deviation $\sigma_{\bar{D}_i}$. MultiSURF uses $\alpha = 1/2$ [21].

Previously we showed empirically for balanced binary outcome datasets that a good constant-k approximation to the expected number of neighbors within the multiSURF radii is $k = m/6$ [7], where m is the number of samples. In Ref. [20] we derive a more exact theoretical mean that shows the mathematical connection between fixed-α and fixed-k neighbor-finding methods, which is given by

$$\bar{k}_{\alpha} = \left\lfloor \frac{m-1}{2} \left(1 - \text{erf} \left(\frac{\alpha}{\sqrt{2}} \right) \right) \right\rfloor, \tag{7}$$

where we apply the floor to ensure the number of neighbors is integer. For data with balanced hits and misses in standard fixed-k Relief, one further divides this formula by 2, and then for multiSURF ($\alpha = 1/2$), we find $\bar{k}_{1/2}^{\text{hit/miss}} = \frac{1}{2}\bar{k}_{1/2} = .154(m-1)$, which is very close to our previous empirical estimate $m/6$. In the current study when we compare multiSURF neighborhood methods with fixed-k neighborhoods, we use $\bar{k}_{1/2}$. Using this $\alpha = 1/2$ value has been shown to give good feature selection performance by balancing power for main effects and interaction effects. However, the best value for α or k is likely data-specific and may be determined through nested cross-validation and other parameter tuning methods [20].

2.2 Nearest-neighbor Projected-Distance Regression (NPDR) with the generalized linear model

2.2.1 Continuous outcomes: linear regression NPDR

Once the neighborhood $\mathcal{N}$ (Eq. 3) is determined by the distance matrix D_{ij} (Eq. 1) and the choice of neighborhood method (e.g., fixed number of neighbors k or adaptive radius), we can compute the NPDR test statistic and P value for the association of an attribute with the phenotype. The NPDR model predictor vector is the attribute's projected distances (e.g., Eq. 2 for numeric attributes) between all pairs of nearest-neighbor instances (Eq. 3). For continuous outcome data (quantitative phenotypes), the NPDR model outcome vector is the numeric difference (Eq. 2 between all nearest neighbors i and j. We find the parameters of the following model that minimize the least-squares error over $\forall (i,j) \in \mathcal{N}$:

$$\mathrm{d}_{ij}^{\text{num}}(y) = \beta_o + \beta_a \mathrm{d}_{ij}(a) + \epsilon_{ij}. \tag{8}$$

The $\mathrm{d}_{ij}^{\text{num}}(y)$ term on the left is the projected distances (diff) between instances i and j for numeric phenotype y (Eq. 2), and ϵ_{ij} is the error term for this random variable. The predictor attribute a may be numeric or categorical, which determines the "type" used in the diff function on the right hand side of Eq.(8). The NPDR test statistic for attribute a is the β_a estimate with one-sided hypotheses

$$\begin{aligned} H_0 &: \beta_a < 0 \\ H_1 &: \beta_a \geq 0. \end{aligned} \tag{9}$$

The β_a can be interpreted as the predicted amount the quantitative outcome changes between a pair of subjects when the projected difference of the attribute value a changes by one unit. The attribute weights in the original RRelief algorithm [11] can be described as a weighted covariance between the attribute neighbor diffs, $\mathrm{d}_{ij}(a)$, and the outcome neighbor diffs, $\mathrm{d}_{ij}^{\text{num}}(y)$. The extra weighting in RRelief is an exponentially decaying function of the rank of the distance between neighbors. Because the NPDR attribute weight, β_a', is a standardized regression coefficient, when the regression contains no covariate term, it can also be written as the correlation between attribute and outcome neighbor diffs:

$$\beta_a' = \mathrm{corr}\left(\mathrm{d}(\mathbf{y}), \mathrm{d}(\mathbf{a})\right). \tag{10}$$

Therefore, unlike RRelief, the NPDR covariance is divided by the variance of the outcome diffs. Thus, there is a similarity between NPDR and RRelief for regression, but, as we show shortly, NPDR provides an improved attribute estimation in a flexible framework for correcting for additional sources of variation (i.e., confounding covariates) as well handling binary outcomes.

5

2.2.2 Continuous outcomes with covariates

Previous Relief-based methods do not include the ability to adjust for covariates. The
regression formalism of NPDR makes adding covariates straightforward. We simply
compute the projected difference values $\mathrm{d}_{ij}(\vec{y}_{\mathrm{covs}})$ for the covariate attribute(s) between
subjects on the neighborhood $(\forall (i,j) \in \mathcal{N})$ and include this as an additional projected
distance term in the regression model:

$$\mathrm{d}_{ij}^{\mathrm{num}}(y) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \vec{\beta}_{\mathrm{covs}}^T \mathrm{d}_{ij}(\vec{y}_{\mathrm{covs}}) + \epsilon_{ij}. \tag{11}$$

The above vector notation can be expanded as

$$\vec{\beta}_{\mathrm{covs}}^T = \left(\beta_{\mathrm{cov}_1}, \beta_{\mathrm{cov}_2}, \ldots, \beta_{\mathrm{cov}_{p_c}} \right) \tag{12}$$

for the regression coefficients of the p_c covariates and

$$\mathrm{d}_{ij}(\vec{y}_{\mathrm{covs}}) = \left(\mathrm{d}_{ij}^{\mathrm{type}_1}(y_{\mathrm{cov}_1}), \mathrm{d}_{ij}^{\mathrm{type}_2}(y_{\mathrm{cov}_2}), \ldots, \mathrm{d}_{ij}^{\mathrm{type}_{p_c}}(y_{\mathrm{cov}_{p_c}}) \right)^T \tag{13}$$

for the projection differences between instances i and j for each of the p_c covariates
with the appropriate projection type for each covariate data type (e.g., numeric or
categorical). The predictor attribute a may be numeric or categorical, which determines
the type used in $\mathrm{d}_{ij}(a)$. The NPDR test statistic is again β_a with alternative hypothesis
$\beta_a \geq 0$ as in Eq. (9).

2.2.3 Dichotomous outcomes with covariates: logistic regression NPDR

We now apply the GLM formalism to enable NPDR to handle dichotomous outcome
data (e.g., case-control phenotype). The STIR method was designed for statistical
testing in dichotomous data, but as we have seen NPDR can also handle continuous
outcomes and adjust for covariates. We model the probability p_{ij}^{miss} that subjects i and
j are in the opposite class (misses) versus the same class (hits) from the neighbor
projected distances with a logit function. We estimate the parameters of the following
model for $\forall (i,j) \in \mathcal{N}$:

$$\mathrm{logit}(p_{ij}^{\mathrm{miss}}) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \epsilon_{ij}, \tag{14}$$

or with covariates:

$$\mathrm{logit}(p_{ij}^{\mathrm{miss}}) = \beta_0 + \beta_a \mathrm{d}_{ij}(a) + \vec{\beta}_{\mathrm{covs}}^T \mathrm{d}_{ij}(\vec{y}_{\mathrm{covs}}) + \epsilon_{ij}, \tag{15}$$

where p_{ij}^{miss} is the probability that subjects i and j have different phenotypes given the
difference in their values for the attribute, a, and given the covariate differences. The
outcome variable that is modeled by probability p_{ij}^{miss} is a binary diff between subjects
for the phenotype (y):

$$\mathrm{d}_{ij}^{\mathrm{miss}}(\vec{y}) = \begin{cases} 0, & y_i == y_j \\ 1, & \text{else.} \end{cases} \tag{16}$$

The β_a importance score can be interpreted in the following way. For a unit increase in
the difference in the value between two neighbors, we predict a change of e^{β_a} in the
odds of the neighbors being in opposite classes. For binary outcome data, we are
specifically interested in the alternative hypothesis that $\beta_a > 0$ because negative β_a
values represent attributes that are irrelevant to classification. Thus, we are interested
in testing one-sided hypotheses

$$\begin{aligned} H_0 &: \beta_a \leq 0 \\ H_1 &: \beta_a > 0. \end{aligned} \tag{17}$$

Nominal outcomes can be analyzed (similar to multi-state Relief-F) with NPDR by
grouping all miss types together. This may be improved by using multinomial regression
in NPDR.

2.2.4 Regularized NPDR

We now propose a regularized NPDR approach that combines all of the attribute difference vectors into one design matrix and constrains the optimization of the coefficients with an elastic net penalty [22] for feature selection. Specifically, we minimize the vector of regression coefficients, $\vec{\beta}_A$ for all attributes $a \in A$, that is $\vec{\beta}_A^T = (\beta_{a_1}, \beta_{a_2}, \ldots, \beta_{a_p})$, subject to the elastic net penalty:

$$
\min_{\beta_o, \vec{\beta}_A} \frac{1}{|\mathcal{N}|} \sum_{i,j \in \mathcal{N}} \mathcal{L}\left(\mathrm{d}_{ij}^{\mathrm{miss}}(y), \beta_0 + \vec{\beta}_A^T \mathrm{d}_{ij}(A)\right)
$$
$$
+ \lambda \left[\alpha ||\vec{\beta}_A||_1 + ||\vec{\beta}_A||_2^2 (1 - \alpha)/2 \right].
$$

$\mathcal{L}$ is the negative log-likelihood for each pair of instances i and j in neighborhood $\mathcal{N}$, and $\mathrm{d}_{ij}(A)$ represents the vector of diffs for fixed i and j for all attributes $a \in A$:

$$
\mathrm{d}_{ij}(A) = \left(\mathrm{d}_{ij}^{\mathrm{type}_1}(a_1), \mathrm{d}_{ij}^{\mathrm{type}_2}(a_2), \ldots, \mathrm{d}_{ij}^{\mathrm{type}_p}(a_p) \right)^T \tag{18}
$$

The elastic-net parameter α mixes the amount of lasso ($\alpha=1$) and ridge ($\alpha=0$) penalty. Our implementation allows any value of α, but we use lasso as the default to choose one predictor (attribute diff vector) when a set of predictors is correlated. The overall penalty strength λ is chosen by cross-validation. For binary outcome, we use binomial link function for the hit/miss projected distances in the likelihood optimization, which leads to elastic net treating positive and negative β_{a_i} coefficients similarly when shrinking. One can simply remove any negative coefficients after shrinkage, but a better approach may be to modify the likelihood to ordinal (one-sided likelihood).

2.3 Properties of NPDR and existing Relief-based methods

Here we summarize the properties and capabilities of standard Relief-based methods and the generalizations STIR and NPDR (Table 1). When there are no covariates for binary outcome data, the STIR (based on a pseudo t-test) and NPDR (based on regression with a logit model) are approximately equivalent given reasonable distribution assumptions. However, by design, the NPDR framework is more flexible and able to handle covariate adjustment and continuous outcomes. In the current notation, the STIR null and alternative hypotheses would be

$$
\begin{aligned}
H_0^{\mathrm{stir}} &: \mu_M(a) - \mu_H(a) \leq 0 \\
H_1^{\mathrm{stir}} &: \mu_M(a) - \mu_H(a) > 0,
\end{aligned} \tag{19}
$$

where

$$
\begin{aligned}
\mu_M(a) &= \bar{M}_a = E\left(\mathrm{d}_{ij}(a) \cdot \left(1 - \mathrm{d}_{ij}^{\mathrm{miss}}(y)\right) \right) \\
\mu_H(a) &= \bar{H}_a = E\left(\mathrm{d}_{ij}(a) \cdot \mathrm{d}_{ij}^{\mathrm{miss}}(y) \right)
\end{aligned} \tag{20}
$$

and the test statistic is a pseudo t-test (see Ref [7]).

For dichotomous outcomes, STIR improves the attribute estimate over Relief weights by incorporating sample variance of the nearest neighbor distances, which enables STIR to estimate statistical significance with the assumptions of a t-test. NPDR assumes intra- and inter-class differences are randomly sampled from one distribution and computes the importance score from a logistic regression. This generalization enables NPDR to have desirable properties and to be applicable to a wider range of problems.

	Standard Relief-based	STIR	NPDR
Importance score (dichotomous)	$\bar{M}_a - \bar{H}_a$	$\dfrac{\bar{M}_a - \bar{H}_a}{S_p(M,H)\sqrt{\frac{1}{|M|} + \frac{1}{|H|}}}$	β_a' coefficient
Score has a null distribution	No	Yes	**Yes**
Supports continuous outcome	Yes	No	**Yes**
Supports covariates	No	No	**Yes**

Table 1. Properties of standard Relief-based methods and generalizations STIR and NPDR. The coefficient β_a' is the NPDR score for logistic (Eq. 14) or linear regression (Eq. 8). The quantity S_p in STIR is the pooled standard deviation of the hit and miss means. Only the score for dichotomous Relief is shown and STIR is restricted to dichotomous outcome [7].

2.4 Real and simulated datasets

2.4.1 Simulation methods

To compare power and false positive performance for multiple feature selection methods, we use the simulation tool from our private Evaporative Cooling (privateEC) software [23] that was designed to simulate realistic main effects, correlations, and interactions found in gene expression or resting-state fMRI correlation data. In the current study, we first simulate main effect data with $m = 200$ subjects (100 cases and 100 controls) and $p = 1000$ real-valued attributes with 10% functional (true positive association with outcome). We choose a sample size consistent with real gene expression data but on the smaller end to demonstrate a more challenging scenario. Likewise, the effect size parameter ($b = 0.8$) was selected to be sufficiently challenging with power approximately 40% [23].

For interactions, we use the differential co-expression network-based simulation tool in privateEC, which is described in Refs. [3, 23]. We first create a co-expression network on an Erdős-Rényi random graph with 0.1 attachment probability, which is the critical value for a giant component. We give connected genes a higher average correlation, approximately $r_{\mathrm{connected}} = 0.8$. This connection correlation induces interaction effect by disrupting the correlation of these connections. Finally, we simulate functional effects on the outcome of functional variables by permuting their values in cases but leaving the controls unpermuted, thereby leading to a final differential correlation network.

All resulting p-values (from STIR and NPDR) are adjusted for multiple testing. Attributes with adjusted p-values less than 0.05 are counted as a positive test (null hypothesis rejected), else the test is negative. We assess the feature-selection performance of each method by averaging the area under the precision-recall curve (auPRC) across 100 replicates of each simulation scenario. Assuming relatively few functional attributes (10% of 1,000 attributes) compared to non-functional ones, the precision and recall measures are robust to imbalanced data and are thus a useful assessment tool of a method's propensity to assign higher scores to the correct functional attributes. The auPRC is also a good comparison tool for methods that, unlike NPDR, do not have a statistical significance threshold. We remark that, even though the auPRC terminology used here is similar to traditional sample classification problems, we instead focus on evaluating the attribute score quality, not accuracy.

2.4.2 RNA-Seq datasets with confounding factors

To test the ability of NPDR to correct for confounding, we use the RNA-Seq study in Ref. [24] that consists of 15,231 genes in 463 MDD cases and 452 controls. Of the 915 subjects, 641 are female and 274 are male. The chi-square between MDD and sex is 25.746 ($p = 3.89e - 7$), and there are 485 genes associated with sex. Thus, there is high risk for confounding effects due to sex differences. We apply NPDR with multiSURF neighborhood and computed importance scores of all genes with and without sex as a covariate to isolate confounding genes. We also apply NPDR to another RNA-Seq expression dataset with a continuous outcome: Pittsburgh Sleep Quality Index (PSQI). Sleep quality was measured for 78 individuals diagnosed with major depressive disorder (MDD) and 79 healthy controls (HC) [25]. Gene expression levels were quantified from reads of 19,968 annotated protein-coding genes and preprocessed with low read-count removal, technical adjustment, and coefficient of variation filtering. As a result, 5,912 RNA-Seq measures were obtained and used as input in the NPDR model to predict sleep quality by PSQI [26]. All P values were adjusted with the Bonferroni procedure.
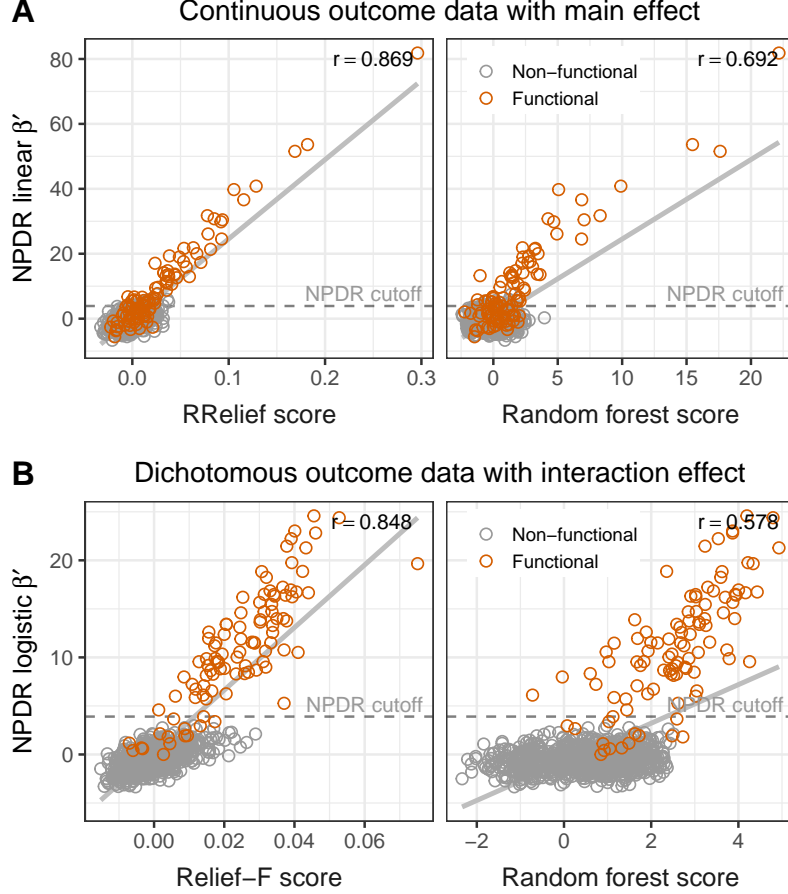
3   Results

3.1   Simulation results

In simulated data with both continuous and dichotomous outcomes, NPDR yields improved attribute estimates than standard Relief and random forest (Fig. 1). In both simulation of main effects with a continuous outcome and interaction effects with a dichotomous outcome, we generate $m = 200$ samples and $p = 1,000$ attributes with 100 functional. In respective simulations, we compare NPDR with a linear model (Eq. 8) and logistic model (Eq. 14) with random forest (permutation importance, R package randomforest) and standard RRelief and Relief-F (R package CORElearn). Because CORElearn does not include the adaptive multiSURF neighborhood, we use a fixed-k neighborhood $\mathcal{N}_{\bar{k}_{1/2}}$ for the Relief-based methods. The value $\bar{k}_{1/2} = 30$ (Eq. 7) is the expected number of nearest neighbors corresponding to a multiSURF neighborhood [20]. With a Bonferroni cutoff, NPDR correctly detects 57 out of 100 functional attributes in the continuous outcome (Fig 1A) and 86 out of 100 functional attributes in the dichotomous outcome data (Fig 1B). As expected, NPDR importance scores are more correlated with Relief-F ($r = 0.869, 0.848$) than with random forest ($r = 0.692, 0.578$).

   Due to the lack of statistical thresholds of random forest and Relief-based methods, in practice, the importance score cutoffs for these methods are chosen arbitrarily. However, when a vertical cutoff is drawn for Relief-based or random forest importance score, it is difficult for random forest and Relief to detect most of the functional attributes without including many false positives, especially for random forest with interaction effects (Fig 1). To explain this underwhelming performance, we recall that random forest underestimates the importance of interacting attributes as the attribute dimensionality becomes too large [27, 28].

   Meanwhile, there is a clearer separation between the functional and non-functional attributes on the NPDR importance metric. This improved feature selection performance of NPDR is better illustrated with the precision-recall curves (Fig 2A). For both datasets with different outcome and effect types, NPDR yields the highest auPRC compared to random forest and Relief-based methods. The auPRC of these methods in the interaction effect simulations are more than main effect simulations, which is due to differences in the simulated effect size. Across 100 replicated simulations for each data type, NPDR shows significantly higher auPRC than random forest and Relief (both P < 0.0001, Fig 2B). Expectedly, when the effect size increases, the auPRC values of all

9

Fig 1. Comparison of NPDR versus Relief-based and random forest importance scores for continuous outcome data with main effects (top row) and dichotomous outcome data with interaction effects (bottom row). Results for one replicate simulation ($m = 200$ samples and $p = 1,000$ attributes with 100 functional). For continuous outcome (A), importance scores computed by RRelief weight, random forest percent increase in MSE and NPDR standardized linear regression coefficient. For dichotomous outcome (B), scores computed by Relief-F, random forest mean decrease in accuracy and NPDR standardized logistic regression coefficient. A regression line between the scores with correlation r is displayed, and a 0.05 Bonferroni cutoff (dashed) is shown for NPDR. There is no statistical threshold for Relief-based methods, so area under the precision-recall curve (auPRC) is used to compare algorithm performance (Fig. 2).
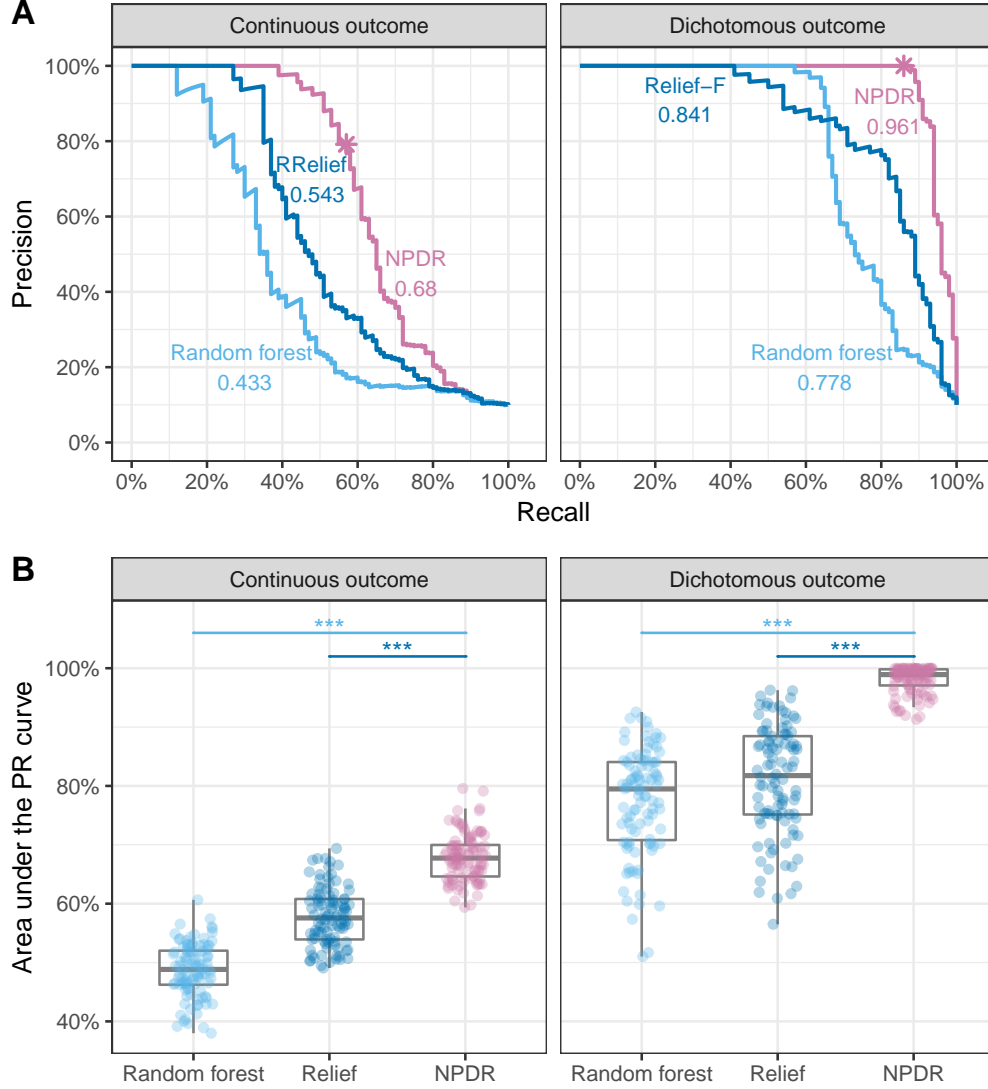
Fig 2. Area under the precision-recall curve (auPRC) of NPDR, random forest and Relief-based importance scores for continuous outcome data with main effects (left) and dichotomous outcome data with interaction effects (right) for one replicate simulation (A) and 100 replicate simulations (B). All simulations use $m = 200$ samples and $p = 1,000$ attributes with 100 functional. The stars $*$ indicate the NPDR cutoffs. NPDR yields significantly higher auPRC in both simulation types.

methods approach 100% and become closer together. Moreover, even though area under the Receiver Operating Characteristics curve (auROC) is not our primary metric, NPDR also gives significantly higher auROC than random forest and Relief (both $P < 0.0001$, see Supplementary Fig S1). In summary, as the effect size varies, NPDR produces distance regression-based importance scores at least as good as if not better than random forest and Relief-based method.

We also compare NPDR (based on logistic regression) and STIR (based on a t-test) for the dichotomous outcome data with interaction effects. While logistic regression and the t-test have slightly different assumptions on the distribution of samples, NPDR and STIR yield highly correlated scores for dichotomous data with interaction effects, where the correlation value r between the P values produced from the two methods ranges from 0.9827 to 0.9994 in 100 replications (Supplementary Fig S2).

For a dataset of the size simulated in our study ($m = 200$ samples and $p = 1000$ attributes), on a desktop with an Intel Xeon W-2104 CPU and 32GB of RAM, NPDR has a 24-second and 3-second runtime for binary and continuous outcome data, respectively.

## 3.2 Real-world RNA-Seq data for MDD with confounding

We apply NPDR with the multiSURF neighborhood $\mathcal{N}_{\alpha=1/2}$ and an adjustment for the sex covariate (Eq. 15) to the RNA-Seq study of MDD in Ref. [24]. This study contains numerous genes that are potentially confounded by sex differences. The NPDR adjustment removes the genes that are most likely spurious associations due to confounding (dark points below the horizontal 0.05 adjust significance line in Fig. 3) compared to NPDR without adjustment. Not only do these removed genes have strong differential expression based on sex, but many of these genes, such as PRKY, UTY, and USP9Y, are Y-linked (i.e., Y chromosome (male)) and mainly expressed in testis. For example, the RPS4Y2 ribosomal protein S4 Y-linked 2 has been shown by tissue specific studies to mainly express in prostate and testis [29] while RPS4X is most expressed in the ovary. In summary, this comparison suggests the NPDR with covariate adjustment effectively removes sex-related confounding variables. A list of 485 genes with expression most associated with sex (Bonferroni-adjusted P value < 0.0001) and their statistical significance are included in the Supplementary Table S1.

The NPDR runtime for this RNA-Seq dataset ($m = 915$ samples and $p = 15,231$ attributes) was approximately 2.3 hours on a desktop with an Intel Xeon W-2104 CPU and 32GB of RAM. As demonstrated in the subsequent section, when typical filtering of the attributes is applied, the runtime is greatly reduced.

3.3 Real-world RNA-Seq data analysis with continuous outcome (sleep quality)

The Pittsburgh Sleep Quality Index (PSQI) measures the quality and patterns of sleep over a one-month interval, where low PSQI indicates high sleep quality. Patients with MDD on average have significantly higher PSQI (9.7) compared to healthy participants (5.0) ($P < 0.0001$, see Supplementary Fig S3). NPDR identifies 24 genes significantly associated with PSQI (Table 2). Among these genes, CACNA1C has been found to be associated with narcolepsy [30], and its genetic variation has been linked to increase risk for sleep latency in infants [31] and psychiatric disorders, including bipolar disorder [32], schizophrenia, recurrent major depression [33] and across psychiatric disorders [34, 35]. Further characterization of the interactions among these genes will prove useful in establishing an sleep-associated expression-epistasis network [3, 27].

Notably, when depression status instead of sleep quality is considered, NPDR with and without sex adjustment produced results that are consistent with previously
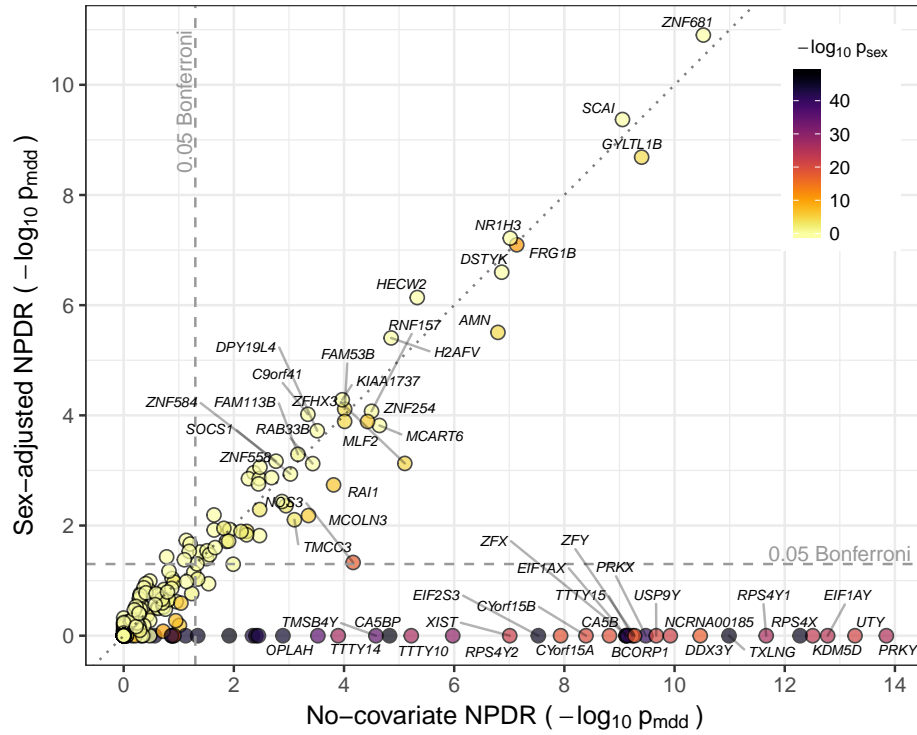
Fig 3. Gene scatter plot of $-\log_{10}$ **significance for association with major depressive disorder** using NPDR without correction for sex (horizontal axis) and with correction for sex (vertical axis). Genes with adjusted $p_{\text{mdd}} < 0.001$ by either method are labeled. NPDR without sex correction finds 87 genes associated with MDD at the Bonferroni-adjusted 0.05 level (right of vertical dashed line), 53 of which are also significantly correlated with sex (adjusted $p_{\text{sex}} < 0.05$). NPDR with adjustment for sex finds 56 genes associated with MDD at the Bonferroni-adjusted 0.05 level (above horizontal dashed line), 19 of which are significantly correlated with sex. The most highly associated genes with sex are eliminated by adjustment (dark genes below the horizontal dashed line) but remain in the non-adjusted set (right of dashed vertical line).

Gene	Description	Beta	P value	P adjusted
ARHGEF4	Rho guanine nucleotide exchange factor 4	6.18	3.39e-10	2.01e-06
C11orf21	Chromosome 11 open reading frame 21	5.31	5.58e-08	3.30e-04
STX4	Syntaxin 4	5.28	6.55e-08	3.87e-04
PTPRE	Protein tyrosine phosphatase, receptor type E	5.19	1.06e-07	6.28e-04
UHRF1	Ubiquitin like with PHD and ring finger domains 1	5.15	1.32e-07	7.78e-04
AGAP2	Arfgap with gtpase domain, ankyrin repeat and PH domain 2	4.90	4.78e-07	0.003
C16orf45	Chromosome 16 open reading frame 45	4.90	4.95e-07	0.003
DPH3P1	Diphthamide biosynthesis 3 pseudogene 1	4.75	1.06e-06	0.006
ACP5	Acid phosphatase 5, tartrate resistant	4.65	1.68e-06	0.010
CBLL1	Cbl proto-oncogene like 1	4.65	1.72e-06	0.010
QRICH1	Glutamine rich 1	4.56	2.59e-06	0.015
RAB7A	Rab7a, member ras oncogene family	4.53	2.93e-06	0.017
PSMD5	Proteasome 26s subunit, non-atpase 5	4.48	3.86e-06	0.023
CBY1	Chibby family member 1, beta catenin antagonist	4.47	4.02e-06	0.024
CACNA1C	Calcium voltage-gated channel subunit alpha1 C	4.46	4.08e-06	0.024
TUBA4B	Tubulin alpha 4B	4.41	5.16e-06	0.031
WNT8B	Wnt family member 8B	4.40	5.59e-06	0.033
PMEL	Premelanosome protein	4.36	6.60e-06	0.039
EFR3B	Efr3 homolog B	4.34	7.36e-06	0.044
MYLPF	Myosin light chain, phosphorylatable, fast skeletal muscle	4.33	7.45e-06	0.044
NSG1	Neuronal vesicle trafficking associated 1	4.33	7.66e-06	0.045
CHKA	Choline kinase alpha	4.33	7.67e-06	0.045

Table 2. Genes associated with Pittsburgh Sleep Quality Index (PSQI) in NPDR analysis.

published results using the STIR algorithm (Supplementary Fig S4) [7]. We expected this consistency because of the balance in sex between diagnostic groups was designed to reduce confounding. The NPDR runtime was 99 seconds (on desktop with an Intel Xeon W-2104 CPU and 32GB of RAM).

4   Discussion

NPDR is the first method to our knowledge to combine projected distances and nearest-neighbors into a generalized linear model regression framework to perform feature selection. The use of nearest neighbors enables its ability to detect interacting attributes, which is shared with Relief-based methods, but NPDR is a departure from Relief in five ways. (1) For feature selection with a continuous outcome, it does not rely on the idea of a hit/miss group like current RRelief approaches [19]. Rather, NPDR simply performs a regression between the outcome and attribute projected distances. (2) For feature selection with binary outcomes, NPDR uses a logistic model to fit pairwise projected distance regressors of hit and miss group. (3) This distance-based regression formalism provides a simple mechanism for NPDR to correct for covariates, which is often neglected in machine learning and has been a limitation of Relief-based methods. (4) For any outcome data type (dichotomous or continuous) and predictor data type, NPDR computes the statistical significance of attribute importance scores, which allows for statistically based thresholds that can adjust for multiple hypothesis testing. (5) The NPDR attribute estimator (regression coefficient) includes more variation from the projected differences than other Relief-based methods and thereby improves attribute estimate quality. Moreover, we introduced a regularized NPDR that adds another layer of multivariate modeling to an already multi-dimensional nearest-neighbor method to

shrink correlated projected attribute differences.

The regression formalism of NPDR is a novel way to perform RRelief by defining the attribute importance as the standardized regression coefficient between the projected attribute differences and the numeric outcome differences between neighbors. For linear regression, NPDR shares some similarity with the original RRelief algorithm because the regression coefficient is related to the correlation coefficient. The RRelief importance score is a weighted correlation between attribute and outcome differences, while the NPDR regression coefficient is a covariance between attribute and outcome differences divided by the variance in the outcome. We showed that the NPDR regression coefficient is a better attribute estimator than RRelief and the NPDR models may include correction terms for covariates that may be confounding and other sources of variation.

The original formulation of Relief focused on binary outcome data, which lends itself naturally to nearest hit (same class) and miss (opposite class) neighbors. As discussed above, for regression problems, the original regression Relief (RRelief) score was cast as a weighted correlation between outcome and the attribute diff [36]. In a different approach, Ref. [19] uses a standard deviation of the continuous outcome diffs to discretize the numeric outcome and make the RRelief algorithm compatible with the idea of hits and misses. However, discretization puts constraints on the variation in numeric data and induces the risk of losing power. NPDR uses the full variation in the continuous outcome variable, and the regression coefficient provides an interpretation in terms of variation explained while again providing flexibility for modeling additional effects.

We assessed NPDR's power and ability to control false positives using realistic simulations with main effects and network interactions. We showed that the statistical performance using NPDR P values is the same as the original STIR, which is specific to dichotomous outcome data. In other words, by modeling hit/miss differences between neighbors with a logit link, NPDR can be safely used instead of STIR with the added benefit of covariate correction and the analysis of quantitative traits.

The incorporation of covariates in NPDR addresses the important but often neglected issue of confounding factors in machine learning. We applied NPDR to a real RNA-Seq dataset for MDD to demonstrate the identification of biologically relevant genes and the removal of spurious associations by covariate correction. NPDR with sex as a covariate adjustment successfully removed X and Y linked genes and genes highly expressed in sex organs. It is important to note that some genes removed due to a shared association with sex may be important for the pathophysiology of MDD or for classifiers. Thus, covariate adjustment in NPDR is a useful option to inform a holistic analysis of a given dataset. Application to GWAS data requires no additional modifications of the algorithm other than specification of a different diff function for categorical variables [6], and the covariate option allows for principal components to be included to adjust for population structure.

[Include? Discuss dependency in NPDR regression? Further improvements in the NPDR covariate adjustment may be achieved by correcting the distance matrix calculation. Adding a covariate term may not be sufficient in some cases because the neighborhood could be strongly influenced by confounding genes due to the curse of dimensionality. The curse of dimensionality can also affect confounding in distance matrix calculation.]

NPDR can use fixed-k Relief neighborhoods and radius-based Relief neighborhoods. For fixed-k neighborhoods, we expect the NPDR approach will handle imbalanced data in a less biased way than the original fixed-k methods, which focus on hit/miss neighborhoods separately. By identifying the nearest neighbors independently of hit/miss status, the neighborhood should naturally reflect the imbalance in the data.

The hit/miss status of each pair is computed separately as a categorical outcome regression variable. This should make NPDR scores with fixed-k ($\mathcal{N}_{k_\alpha}$) similar to fixed radius ($\mathcal{N}_{R_\alpha}$) for balanced and imbalanced data.

A related distance-based regression method is Multivariate Distance Matrix Regression (MDMR) [37]. The MDMR approach uses an F-statistic to test for the association of distance matrices between two sets of factors. The MDMR regression is performed for the distance matrix for all pairs of instances, not a subset of nearest neighbors, which makes it susceptible to missing interactions. NPDR projects distances onto each attribute, allowing for hypothesis testing of individual attributes (i.e., perform feature selection), whereas MDMR focuses on sets of attributes. While NPDR uses the context of all attributes to compute nearest neighbors, it focuses on the projected regression of each attribute at a time and uses the nearest neighbors to allow for detection of interactions. MDMR uses all pairs of subjects in the regression, making it more myopic and susceptible to missing interaction effects. The ability to remove imposters from the set of nearest neighbors illustrates the blessings of dimensionality for Relief-based methods [20], but this class of nearest-neighbor methods is still, of course, susceptible to the curses of dimensionality [38]. NPDR can also be used to compute the importance of sets of factors. An example of this is the penalized version of NPDR that uses the set of all attributes in a multiple nearest-neighbor projected-distance regression.

Power for detecting main effects is highest with the myopic maximum $k = k_{\max} = \lfloor (m-1)/2 \rfloor$. Real biological data will likely contain a mixture of main effects and epistasis network effects [39]. STIR feature selection could be embedded in the backwards elimination of private Evaporative Cooling (privateEC) for feature selection and classification [23] or embedded in a nested cross-validation approach. Nested CV and privateEC can also return classification and optimize α or k.

Acknowledgements

Funding

References

1. MS Breen, C Kemena, PK Vlasov, C Notredame, and Kondrashov FA. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–8, 2012.

2. DM Weinreich, Y Lan, e CS Wyli, and RB Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev.*, 23(6):700–7, 2013.

3. Caleb A Lareau, Bill C White, Ann L Oberg, and Brett A McKinney. Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure. *BioData mining*, 8(1):5, 2015.

4. Alberto De la Fuente. From differential expression to differential networking–identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7):326–333, 2010.

5. Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

6. Marziyeh Arabnejad, BD Dawkins, WS Bush, BC White, AR Harkness, and Brett A McKinney. Transition-transversion encoding and genetic relationship metric in relieff feature selection improves pathway enrichment in gwas. *BioData mining*, 11:23, 2015.

7. Trang T Le, Ryan J Urbanowicz, Jason H Moore, and Brett A McKinney. Statistical inference relief (stir) feature selection. *Bioinformatics*, 2018.

8. Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 2018.

9. Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, January 1997.

10. Brett A. McKinney, James E. Crowe, Jingyu Guo, and Dehua Tian. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS genetics*, 5(3):e1000432, March 2009.

11. Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

12. Brett A McKinney, Bill C White, Diane E Grill, Peter W Li, Richard B Kennedy, Gregory A Poland, and Ann L Oberg. Reliefseq: a gene-wise adaptive-k nearest-neighbor feature selection tool for finding gene-gene interactions and main effects in mrna-seq gene expression data. *PloS one*, 8(12):e81527, 2013.

13. Trang T Le, Rayus T Kuplicki, Brett A McKinney, Hung-wen Yeh, Wesley K Thompson, and Martin P Paulus. A nonlinear simulation framework supports adjusting for age when analyzing brainage. *Frontiers in aging neuroscience*, 10, 2018.

14. Han Chen, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, Wei Chen, John M Brehm, Juan C Celedón, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666, 2016.

15. Anil Rao, Joao M Monteiro, Janaina Mourao-Miranda, Alzheimer's Disease Initiative, et al. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*, 150:23–49, 2017.

16. Kristin A Linn, Bilwaj Gaonkar, Jimit Doshi, Christos Davatzikos, and Russell T Shinohara. Addressing confounding in predictive models with an application to neuroimaging. *The international journal of biostatistics*, 12(1):31–44, 2016.

17. Limin Li, Barbara Rakitsch, and Karsten Borgwardt. ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics*, 27(13):i342–i348, 2011.

18. Casey S. Greene, Nadia M. Penrod, Jeff Kiralis, and Jason H. Moore. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining*, 2:5, September 2009.

19. Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, 85:168–188, 2018.

20. Bryan A Dawkins, Trang T Le, and Brett A McKinney. Blessings of dimensionality: Theoretical analysis of nearest-neighbor projected-distance methods for detecting interactions in high dimension. *Under Construction*, 2019.

21. Delaney Granizo-Mackenzie and Jason H Moore. Multiple threshold spatially uniform relieff for the genetic analysis of complex human diseases. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 1–10. Springer, 2013.

22. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

23. Trang T Le, W Kyle Simmons, Masaya Misaki, Jerzy Bodurka, Bill C White, Jonathan Savitz, and Brett A McKinney. Differential privacy-based evaporative cooling feature selection and classification with relief-f and random forests. *Bioinformatics*, 33(18):2906–2913, 2017.

24. Sara Mostafavi, Alexis Battle, Xiaowei Zhu, James B Potash, Myrna M Weissman, Jianxin Shi, Kenneth Beckman, Christian Haudenschild, Courtney McCormick, Rui Mei, et al. Type i interferon signaling genes in recurrent major depression: increased expression detected by whole-blood rna sequencing. *Molecular psychiatry*, 19(12):1267, 2014.

25. Trang T. Le, Jonathan Savitz, Hideo Suzuki, Masaya Misaki, T. Kent Teague, Bill C. White, Julie H. Marino, Graham Wiley, Patrick M. Gaffney, Wayne C. Drevets, Brett A. McKinney, and Jerzy Bodurka. Identification and replication of RNA-Seq gene network modules associated with depression severity. *Translational Psychiatry*, 8(1):180, September 2018.

26. Daniel J. Buysse, Charles F. Reynolds, Timothy H. Monk, Susan R. Berman, and David J. Kupfer. The pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*, 28(2):193 – 213, 1989.

27. Brett A McKinney, James E Crowe Jr, Jingyu Guo, and Dehua Tian. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS genetics*, 5(3):e1000432, 2009.

28. Stacey J Winham, Colin L Colby, Robert R Freimuth, Xin Wang, Mariza De Andrade, Marianne Huebner, and Joanna M Biernacka. Snp interaction detection with random forests in high-dimensional genetic data. *BMC bioinformatics*, 13(1):164, 2012.

29. Alexandra M Lopes, Ricardo N Miguel, Carole A Sargent, Peter J Ellis, António Amorim, and Nabeel A Affara. The human rps4 paralogue on yq11. 223 encodes a structurally conserved ribosomal protein and is preferentially expressed during spermatogenesis. *BMC molecular biology*, 11(1):33, 2010.

30. Mihoko Shimada, Taku Miyagawa, Minae Kawashima, Susumu Tanaka, Yutaka Honda, Makoto Honda, and Katsushi Tokunaga. An approach based on a genome-wide association study reveals candidate loci for narcolepsy. *Human Genetics*, 128(4):433–441, Oct 2010.

31. Katri Kantojärvi, Johanna Liuhanen, Outi Saarenpää-Heikkilä, Anna-Liisa   584
    Satomaa, Anneli Kylliäinen, Pirjo Pölkki, Julia Jaatela, Auli Toivola, Lili Milani,   585
    Sari-Leena Himanen, Tarja Porkka-Heiskanen, Juulia Paavonen, and Tiina Paunio.   586
    Variants in calcium voltage-gated channel subunit alpha1 c-gene (cacna1c) are   587
    associated with sleep latency in infants. *PLOS ONE*, 12(8):1–12, 08 2017.   588

32. Barbara Franke, Alejandro Arias Vasquez, Joris A. Veltman, Han G. Brunner,   589
    Mark Rijpkema, and Guillén Fernández. Genetic variation in cacna1c, a gene   590
    associated with bipolar disorder, influences brainstem rather than gray matter   591
    volume in healthy individuals. *Biological Psychiatry*, 68(6):586 – 588, 2010.   592
    Animal Models for Mania and Melancholia.   593

33. E K Green, D Grozeva, I Jones, L Jones, G Kirov, S Caesar, K Gordon-Smith,   594
    C Fraser, L Forty, E Russell, M L Hamshere, V Moskvina, I Nikolov, A Farmer,   595
    P McGuffin, Wellcome Trust Case Control Consortium, P A Holmans, M J Owen,   596
    M C O'Donovan, and N Craddock. The bipolar disorder risk allele at cacna1c   597
    also confers risk of recurrent major depression and of schizophrenia. *Molecular*   598
    Psychiatry, 15:1016 EP –, 07 2009.   599

34. Identification of risk loci with shared effects on five major psychiatric disorders: a 600
    genome-wide analysis. *The Lancet*, 381(9875):1371 – 1379, 2013.   601

35. Amit Lotan, Michaela Fenckova, Janita Bralten, Aet Alttoa, Luanna Dixson,   602
    Robert W. Williams, and Monique van der Voet. Neuroinformatic analyses of   603
    common and distinct genetic components associated with major neuropsychiatric   604
    disorders. *Frontiers in Neuroscience*, 8:331, 2014.   605

36. Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of   606
    relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.   607

37. Nicholas J Schork and Matthew A Zapala. Statistical properties of multivariate   608
    distance matrix regression for high-dimensional data analysis. *Frontiers in*   609
    genetics, 3:190, 2012.   610

38. Naomi S. Altman and Martin Krzywinski. The curse(s) of dimensionality   611
    this-month. *Nature Methods*, 15(6):399–400, 6 2018.   612

39. Brett McKinney and Nicholas Pajewski. Six degrees of epistasis: statistical   613
    network models for gwas. *Frontiers in genetics*, 2:109, 2012.   614