Blessings of Dimensionality: Theoretical analysis of nearest-neighbor projected-distance methods for detecting interactions in high dimension

Bryan A. Dawkins[1], Trang T. Le[2] and Brett A. McKinney[1,3,*]
[1]Department of Mathematics, University of Tulsa, Tulsa, OK 74104, USA
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104
[3]Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104, USA.

Abstract

It is commonly known that high-throughput data has many inherent statistical challenges, such as multiple testing, sparsity and over fitting. Collectively these challenges are known as the Curse of Dimensionality. Here we highlight an important Blessing of Dimensionality: the ability to identify interactions with nearest neighborhoods. We review nearest-neighbor concepts for finding interactions, and we derive important distribution moments for distance metrics in high dimensional spaces. We use these theoretical results and simulated data to offer recommendations for computational approaches to find nearest neighbors in high dimension. We discuss ways to maximize the blessings and minimize the curses of dimensionality to reliably identify interactions.

Author summary

Introduction

Relief-based methods identify interacting attributes as important by using nearest-neighbor information in higher dimensions (the "blessings of dimensionality"). Myopic methods, such as univariate tests, that do not account for information from higher dimensions, are susceptible to false negatives when there are interactions. For example in the plot of variable A versus C in a three-variable simulation (Fig. 1a), variable A appears to show no difference between cases and controls (the marginal group means are the same). However, A is actually simulated to have a strong differential correlation with B, conditioned on the outcome variable (Fig. 2b). Current Relief-based methods determine the importance of an attribute by computing the average difference of a target instance (X) and its nearest instance form the same class (Hit) projected onto the attribute A dimension ($d_{\mathrm{X,H}}(A)$) subtracted from the projected difference of target X and its nearest instance from the opposite class (Miss) ($d_{\mathrm{X,M}}(A)$). When the inequality $d_{\mathrm{X,M}}(A) > d_{\mathrm{X,H}}(A)$, it suggests that attribute A is useful for discriminating between cases and controls.

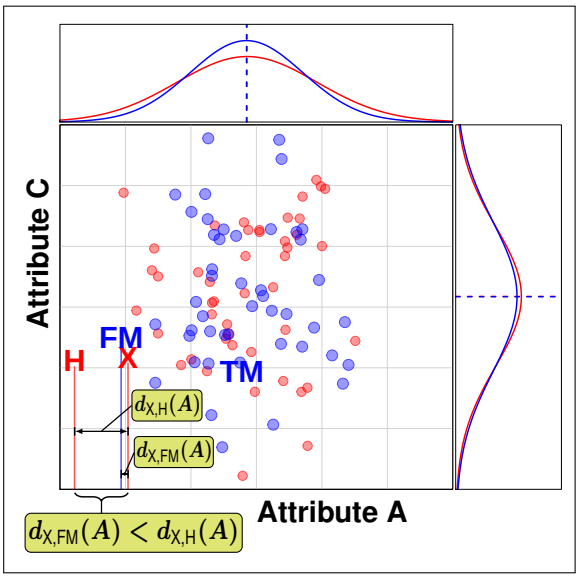

Fig 1. Imposters vs true neighbors in the presence of interactions with three variables. Scatter plot of simulated irrelevant Attribute C with a functional Attribute A **(a)**. None of the attributes has a main effect, but Attribute B and C interact through differential correlation **(b)**. Computing nearest neighbors with irrelevant attributes **(a)** or lower dimensions leads to imposter nearest neighbors and degrades the ability of Relief-based methods to identify interaction effects. Computing distances in only these two dimensions leads to an imposter false miss (FM) for the nearest neighbor from the opposite outcome class for target instance X. This imposter leads to attribute A predicting closer projected distances for misses than hits (H), which incorrectly indicates that A is a poor discriminator (yellow boxes in **(a)**). Computing nearest neighbors in higher dimensions **(c-d)** or with the correct interaction partner leads to imposter nearest neighbor (FM) being replaced by the true nearest miss neighbor (TM) for target instance X, which correctly leads to attribute A predicting closer projected distances for hits (H) than misses, which is an indication that attribute A is a good discriminator (yellow boxes **(b)**).

Relief-based methods use information from all attributes available to it (omnigenic) to estimate an attribute's importance. However, if relevant higher-dimensional information is not used, even Relief-based methods will miss the effect of A because "imposter" neighbors will be used in the attribute estimate (False Miss (FM) in Fig. 1, where $d_{\mathrm{X,FM}}(A) < d_{\mathrm{X,H}}(A)$). If one were to compute nearest neighbors in the A-C plane (ignoring the B dimension), the nearest miss would be an imposter (FM), which leads to a negative contribution to the importance score for A. One might call this C attribute a type-I confounding attribute because it increases the chances of interacting attributes to be false negatives. When nearest neighbors are calculated based on higher dimensions with relevant information (Fig. 2c), it is clear that TM is closer to X than FM. The imposter (FM) is replaced by the true nearest miss (TM) and attribute A correctly shows a greater projected difference between misses than hits (Fig. 2d $d_{\mathrm{X,TM}}(A) > d_{\mathrm{X,H}}(A)$), which is the signature of an important attribute. Univariate methods still cannot find the importance of A unless the interaction is explicitly modeled, but as long as functional variables A and B are in the space for nearest neighbor calculations (Fig. 2c-d), imposters can be excluded and Relief-based methods will find that A (and B) are important discriminators.

[Ideas: Relating the increasing k and myopic view to other distance-related method such as MDS/t-SNE vs PCA - local vs global distance - capturing non-linear manifold structure.
`https://www.kdnuggets.com/2018/08/introduction-t-sne-python.html`]

Using same interaction, increase background noise genes to see degrading of A and B Relief importance because of curse of dimensionality (sparseness).

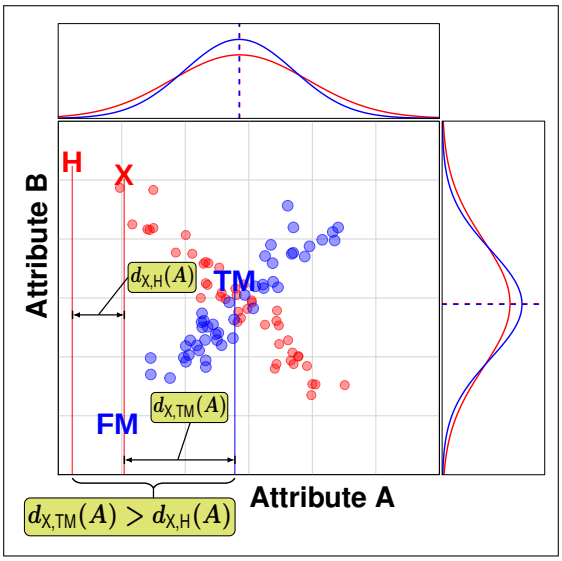

Fig 2. True neighbors

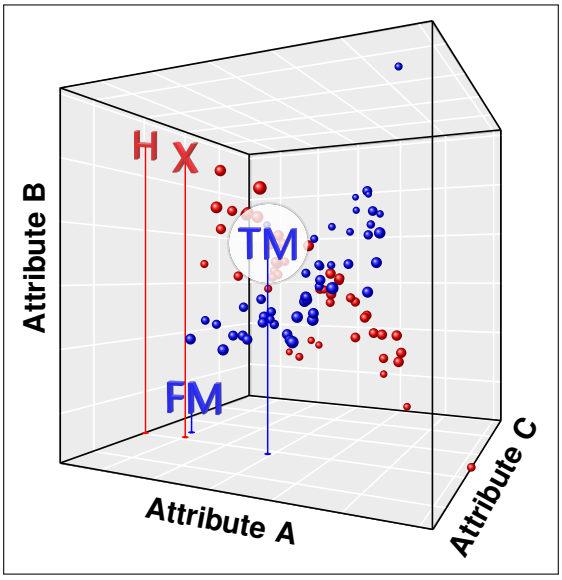

Fig 3. 3D AB view. Still working on this.

1 Neighborhood Methods

NPD methods rely on a neighborhood algorithm for feature selection. One may specify a fixed-k number of neighbors, an average radius SURF, a multiSURF radius that adapts for each instance [1], or a gene-wise adaptive-k.

2 Derivation of expected k for multiSURF neighborhoods

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. Previously we showed

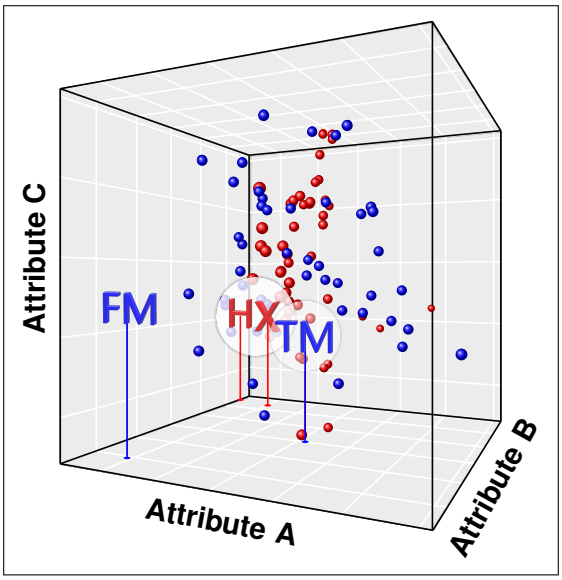

Fig 4. 3D AC view. Still working on this.

empirically for balanced case-control datasets that a good constant-k approximation to the expected number of neighbors within the multiSURF radii is $k = m/6$ [2], where m is the number of samples. Here we derive a more exact theoretical mean that shows the mathematical connection between neighbor-finding methods. This fixed-k approximation to multi-SURF is independent of the type of data and the particular radii of each instance in the data.

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

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

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

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

where $\hat{X}$ represents the standardized data matrix X. We use a simplified $\mathrm{d}_{ij}(a)$ notation in place of the $\mathrm{diff}(a,(i,j))$ notation that is customary in Relief-based methods. We omit the division by $\max(a) - \min(a)$ used by Relief to constrain scores to the interval from -1 to 1. As we show in subsequent sections, NPDR scores are [standardized] regression coefficients with corresponding P values, so any scaling operation at this stage is unnecessary for comparing attribute scores. The numeric $\mathrm{d}_{ij}^{\mathrm{num}}(a)$ projection is simply the absolute difference between row elements i and j of the data matrix $X^{m\times p}$ for the attribute column a.

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 Distribution of pairwise distances

Suppose that $X_{ia}, X_{ja} \overset{iid}{\sim} \mathcal{F}_X(\mu_X, \sigma_X^2)$ for two fixed and distinct instances $i, j \in \{n\}_{n=1}^{m}$ and a fixed attribute $a \in \mathcal{A}$.

It is clear that the magnitude difference, or numeric diff given by Eq. 2, between X_{ia} and X_{ja} is another random variable Z_a. That is,

$$Z_a = d_{ij}(a) = |X_{ia} - X_{ja}| \sim \mathcal{F}_Z(\mu_z, \sigma_z^2). \tag{4}$$

Furthermore, the collection of all magnitude differences between fixed instances i and j is a random sample of size p from $\mathcal{F}_Z(\mu_z, \sigma_z^2)$. That is,
$Z_{a_1}, Z_{a_2}, \ldots, Z_{a_p} \overset{iid}{\sim} \mathcal{F}_Z(\mu_z, \sigma_Z^2)$ such that

$$Z_{a_k} = d_{ij}(a_k) = |X_{ia_k} - X_{ja_k}|. \tag{5}$$

Therefore, the sum of all Z_{a_k} for $k = 1, 2, \ldots, p$ is asymptotically normal by the Classical Central Limit Theorem (CCLT). Explicitly, this means that

$$D_{ij}^{(1)} = \sum_{k=1}^{p} |X_{ia_k} - X_{ja_k}| = \sum_{k=1}^{p} d_{ij}(a_k) = \sum_{k=1}^{p} Z_{a_k} \dot{\sim} \mathcal{N}(\mu_z p, \sigma_z^2 p). \tag{6}$$

Therefore, the standard manhattan distance between two distinct and fixed instances i and j is asymptotically normal.

It is also apparent that the squared difference, or squared numeric diff, between X_{ia} and X_{ja} is another random variable Z_a^2. That is,

$$Z_a^2 = d_{ij}^2(a) = (X_{ia} - X_{ja})^2 \sim \mathcal{F}_{Z^2}(\mu_{z^2}, \sigma_{z^2}^2). \tag{7}$$

Furthermore, the collection of all squared differences between fixed instances i and j is a random sample of size p from $\mathcal{F}_{Z^2}(\mu_{z^2}, \sigma_{z^2}^2)$. That is,
$Z_{a_1}^2, Z_{a_2}^2, \ldots, Z_{a_p}^2 \overset{iid}{\sim} \mathcal{F}_{Z^2}(\mu_{z^2}, \sigma_{z^2}^2)$ such that

$$Z_{a_k}^2 = (X_{ia_k} - X_{ja_k})^2. \tag{8}$$

Therefore, the sum of all $Z_{a_k}^2$ for $k = 1, 2, \ldots, p$ is asymptotically normal by the Classical Central Limit Theorem. Explicitly, this means that

$$\left(D_{ij}^{(2)}\right)^2 = \sum_{k=1}^{p} (X_{ia_k} - X_{ja_k})^2 = \sum_{k=1}^{p} d_{ij}^2(a_k) = \sum_{k=1}^{p} Z_{a_k}^2 \dot{\sim} \mathcal{N}(\mu_{z^2} p, \sigma_{z^2}^2 p). \tag{9}$$

Consider the smooth functional transformation $g(\alpha) = \sqrt{\alpha}$ applied to the random variable given by Eq. 9. The Delta Method (cite textbook) can be applied to show that

$$g\left(\left(D_{ij}^{(2)}\right)^2\right) = g\left(\sum_{k=1}^{p} Z_{a_k}^2\right) = \sqrt{\sum_{k=1}^{p}(X_{ia_k} - X_{ja_k})^2} \dot{\sim} \mathcal{N}\left(g(\mu_{z^2}p), [g\prime(\mu_{z^2}p)]^2\sigma_{z^2}^2\right) \tag{10}$$

Therefore, the standard Euclidean distance between two fixed and distinct instances i and j is asymptotically normal.

We have considered the distance metric given by Eq. 1 for $q = 1$ (Manhattan) and $q = 2$ (Euclidean). Without loss of generality, the argument presented previously for $q = 2$ can be used to show asymptotic normality of $D_{ij}^{(q)}$ with $q > 2$. Therefore, the pairwise distance distribution of $D_{ij}^{(q)}$ is asymptotically normal with the assumption of independently and identically distributed instances. This result is true for any null data distribution, whether continuous or discrete.

For distance based learning methods, all pairwise distances are used to determine relative importances for attributes. The collection of all distances above the diagonal in an $m \times m$ distance matrix does not satisfy the independence assumption used in the previous derivations. This is because of the redundancy that is inherent to the distance matrix calcuation. However, this collection is still asymptotically normal with mean and variance approximately equal to those given in Eqs. 6 or 10. Hence, all fixed-radius methods will use a fixed radius that is some fraction of the expected pairwise distance for a given metric and data type. This implies that the probability of a fixed instance j being within a fixed radius of a given instance i can be parameterized by the expected pairwise distance and the variance of the pairwise distance. This probability is obtained by evaluating the normal cumulative distribution function (CDF), with corresponding mean and variance, at the quantile given by the fixed radius. Therefore, we can derive the expected number of neighbors in the neighborhood of a fixed instance i. In other words, for sufficiently large data sets, the sample mean of the number of neighbors in a given neighborhood is well approximated by the product between the total number of possible neighbors and the expected probability of an instance being in a given neighborhood. The total number of possible neighbors for a fixed instance i is always $m - 1$, but this becomes approximately $\lfloor \frac{m-1}{2} \rfloor$ when delineating between possible hits and misses for balanced data.

2.2 Predicted number of neighbors in the multiSURF alpha neighborhood

Regardless of the predictor data type (numeric or categorical), the distribution of the p predictors (uniform, Gaussian, or binomial), or the metric used to compute distances (Manhattan or Euclidean), the $m(m-1)/2$ pairwise distances in the p-dimensional space are well approximated by a normal distribution. An instance j is in the adaptive α-radius neighborhood of i ($j \in N_i^\alpha$) under the condition

$$D_{ij} \leq R_i^\alpha \implies j \in N_i^\alpha, \tag{11}$$

where the threshold radius for instance i is

$$R_i^\alpha = \bar{D}_i - \alpha\, \sigma_{\bar{D}_i} \tag{12}$$

and

$$\bar{D}_i = \frac{1}{m-1} \sum_{j \neq i} D_{ij}^{(\cdot)} \tag{13}$$

is the average of instance i's pairwise distances (using Eq. D Equation) with standard deviation $\sigma_{\bar{D}_i}$. MultiSURF uses $\alpha = 1/2$ [3].

The probability of the remaining $m-1$ instances being inside the α-radius of instance i (R_i^α) can be viewed as $m-1$ Bernoulli trials each with a probability of success q_α. Then the average average number of neighbors is given by

$$\bar{k}_\alpha = (m-1)q_\alpha, \tag{14}$$

from the mean of a binomial random variable. To calculate q_α, we assume the distribution of distances ($\{D_{ij}\}_{j\neq i}$) of neighbors of instance i is normal $N(\bar{D}_i, \sigma_{\bar{D}_i})$. Our empirical studies confirm a normal distribution and that it is robust to data type and metric. Extreme violations of independence of attributes (extreme correlations or interactions) will cause the distribution to be right skewed, but this effect is difficult to observe in real data. Thus, for a Gaussian pairwise distance distribution, the probability q_α for one instance $j \neq i$ to be in the neighborhood of i ($j \in N_i^\alpha$) is given by the area under the mean-centered ($\bar{D}_i$) Gaussian from $-\infty$ to R_i^α. **show Gaussian plot illustration?** This integral can be written in terms of the error function (erf):

$$q_\alpha = \frac{1}{2}\left(1 - \text{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right). \tag{15}$$

And finally using Eqs. (14 and 15) we find

$$\bar{k}_\alpha = \left\lfloor \frac{m-1}{2}\left(1 - \text{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right)\right\rfloor, \tag{16}$$

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 divides this formula by 2. For multiSURF ($\alpha = 1/2$), this formula gives $\bar{k}_{1/2}^{\text{hit/miss}} = \frac{1}{2}\bar{k}_{1/2} = .154(m-1)$, which is very close to our previous empirical estimate $m/6$. 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 performance for simulated data sets. However, the best value for α is likely data-specific and may be determined through nested cross-validation and other parameter tuning methods.

3 Derivation of means and standard deviations for metrics and data distributions

Table 1. Summary of asymptotic distance distributions for common data types. Metrics with subscripts M and E represent Manhattan and Euclidean, respectively. Metrics with superscript * represent a deviation from the standard metric by attribute range normalization. The function $\Phi^{-1}(x)$ denotes the standard normal quantile function, where $x \in (0, 1)$.

Type	Mean	Variance
$\mathcal{N}(0,1) - \mathbf{d_M}$	$\frac{2p}{\sqrt{\pi}}$	$\frac{2p(\pi-2)}{\pi}$
$\mathcal{N}(0,1) - \mathbf{d_M^*}$	$\boxed{\frac{p}{\sqrt{\pi}\mu(m)}}$ where $\mu(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$	$\boxed{\frac{p(\pi-2)}{2\pi\mu^2(m)}}$ where $\mu(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
$\mathcal{N}(0,1) - \mathbf{d_E}$	$\sqrt{2p-1}$	1
$\mathcal{N}(0,1) - \mathbf{d_E^*}$	$\boxed{\frac{\sqrt{2p-1}}{2\mu(m)}}$ where $\mu(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$	$\boxed{\frac{2\log(m)}{\pi^2 + 12\mu^2(m)\log(m)}}$ where $\mu(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
$\mathcal{U}(0,1) - \mathbf{d_M}$	$\frac{p}{3}$	$\frac{p}{18}$
$\mathcal{U}(0,1) - \mathbf{d_M^*}$	$\frac{(m+1)p}{3(m-1)}$	$\frac{(m^3-18m^2-5m+2)p}{18(m^3+m^2+2)(m-1)^2}$
$\mathcal{U}(0,1) - \mathbf{d_E}$	$\sqrt{\frac{p}{6} - \frac{7}{120}}$	$\frac{7}{120}$
$\mathcal{U}(0,1) - \mathbf{d_E^*}$	$\sqrt{\frac{p}{6} - \frac{7}{120}}\left(\frac{m+1}{m-1}\right)$	$\frac{7(m+1)^2(m+2)}{120(m^3+m^2+2)}$

Table 2. Summary of asymptotic distance distributions for rs-fMRI and GWAS data. Metrics with superscript * represent a deviation from the standard metric by attribute range normalization. The function $\Phi^{-1}(x)$ denotes the standard normal quantile function, where $x \in (0, 1)$.

Type	Mean	Variance
rs-fMRI ($\mathbf{d_{ROI}}$)	$\frac{2p(p-1)}{\sqrt{\pi(p-3)}}$	$\frac{4(\pi-2)p(p-1)}{\pi(p-3)}$
rs-fMRI ($\mathbf{d^*_{ROI}}$)	$\boxed{\frac{2p(p-1)}{\mu(m,p)\sqrt{\pi(p-3)}}}$ where $\mu(m,p) = \frac{1}{\sqrt{p-3}}\Phi^{-1}\left(1-\frac{1}{m(p-1)}\right)$	$\boxed{\frac{2[6(p-3)\mu^2(m,p)\log[m(p-1)](\pi-2)-\pi^2]p(p-1)}{\pi(p-3)\mu^2(m,p)(\pi^2+12(p-3)\mu^2(m,p)\log[m(p-1)])}}$ where $\mu(m,p) = \frac{1}{\sqrt{p-3}}\Phi^{-1}\left(1-\frac{1}{m(p-1)}\right)$
GWAS ($\mathbf{d_{GM}}$)	$\boxed{2\sum_{a=1}^{p} F(a)}$ where $F(a) = \left[2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2\right]$, and f_a is the probability of a minor allele at locus a.	$\boxed{2\sum_{a=1}^{p} F(a)[1-2F(a)]}$ where $F(a) = \left[2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2\right]$, and f_a is the probability of a minor allele at locus a.
GWAS ($\mathbf{d_{AM}}$)	$\boxed{2\sum_{a=1}^{p} F(a)}$ where $F(a) = \left[(1-f_a)^3 f_a + f_a^3(1-f_a) + (1-f_a)^2 f_a^2\right]$, and f_a is the probability of a minor allele at locus a.	$\boxed{\sum_{a=1}^{p}\left[G(a) - 4F^2(a)\right]}$ where $F(a) = \left[(1-f_a)^3 f_a + f_a^3(1-f_a) + f_a^3(1-f_a) + (1-f_a)^2 f_a^2\right]$, $G(a) = \left[(1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2\right]$, and f_a is the probability of a minor allele at locus a.
GWAS ($\mathbf{d_{TiTv}}$)	$\boxed{(\gamma_0+\gamma_2+2\gamma_1)\sum_{a=1}^{p} F(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a=1}^{p} G(a)}$ where $F(a) = \left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right]$ and $G(a) = (1-f_a)^2 f_a^2$, f_a is the probability of a minor allele at locus a, and γ_0, γ_1, and γ_2 are probabilities of PuPu, PuPy, and PyPy, respectively, at locus a.	$\boxed{\left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a=1}^{p} F(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a=1}^{p} G(a) + \sum_{a=1}^{p}\left[(\gamma_0+\gamma_2+2\gamma_1)F(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G(a)\right]^2}$ where $F(a) = \left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right]$ and $G(a) = (1-f_a)^2 f_a^2$, f_a is the probability of a minor allele at locus a, and γ_0, γ_1, and γ_2 are probabilities of PuPu, PuPy, and PyPy, respectively, at locus a.

4 Optimal neighborhood parameters for detecting effects

k or α. Balancing blessing and curse of dimensionality.

5 ICA?

Using same interaction, increase background noise genes to see degrading of A and B Relief importance because of curse of dimensionality (sparseness).

References

1. 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.

2. Trang T Le, Ryan J Urbanowicz, Jason H Moore, and Brett A McKinney. Statistical inference relief (stir) feature selection. *Bioinformatics*, page bty788, 2018.

3. 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.