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

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 neighborhoods of instances. We review the nearest-neighbor concept for finding interactions among attributes and 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 that do not account for information from higher dimensions such as univariate tests 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 attribute A value between a target instance (X) and its nearest instance from the opposite class (Miss), $d_{\mathrm{X,M}}(A)$, subtracted from the similarly projected difference of target X and its nearest instance from the same class (Hit), $d_{\mathrm{X,H}}(A)$. A positive value from this calculation, i.e., $d_{\mathrm{X,M}}(A) - d_{\mathrm{X,H}}(A) > 0$, suggests that attribute A is useful for discriminating between cases and controls.

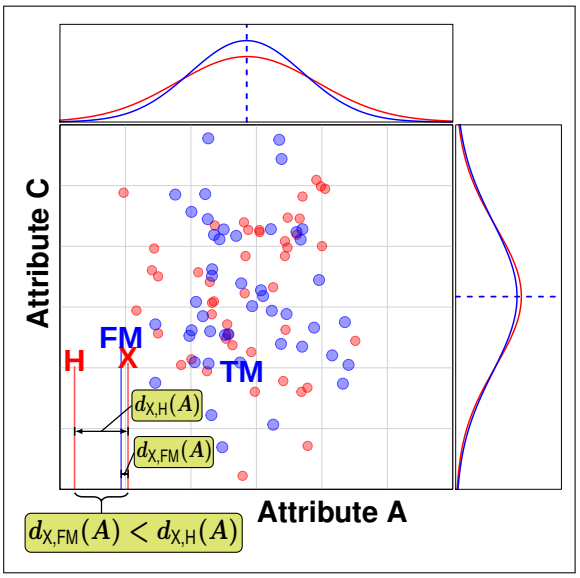

Fig 1. Imposters vs true neighbors in the presence of interactions with three variables. Scatter plot of simulated irrelevant Attribute C with a functional Attribute A **(a)**. None of the attributes has a main effect, but Attribute B and C interact through differential correlation **(b)**. Computing 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 available attributes (omnigenic) to estimate an attribute's importance. However, if relevant higher-dimensional information is not used to establish the neighborhoods of instances, these 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.

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

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

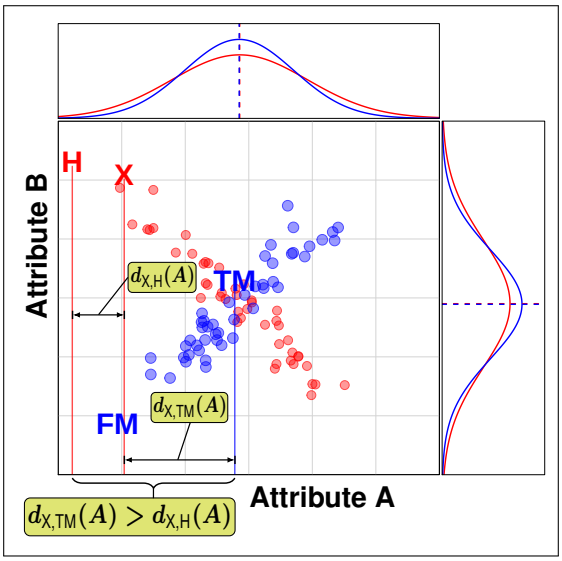

Fig 2. True neighbors

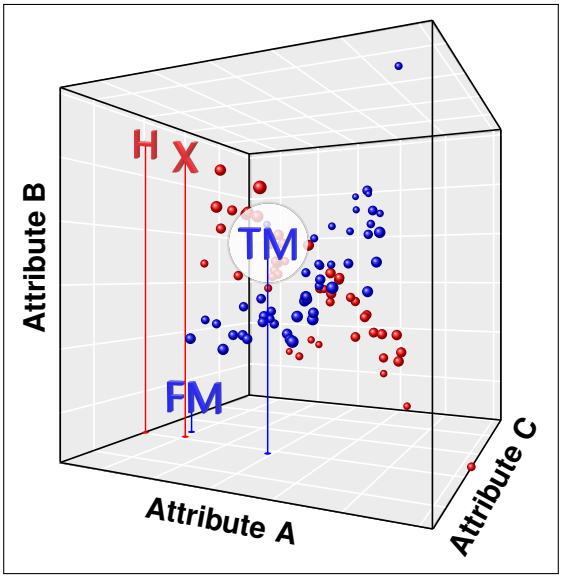

Fig 3. 3D AB view. Still working on this.

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 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

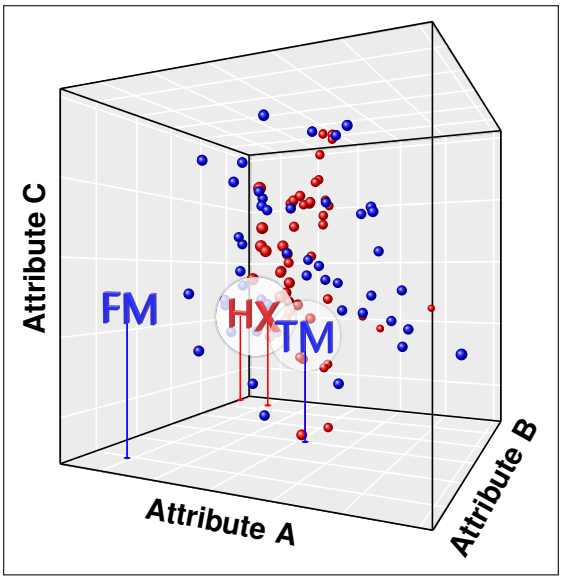

Fig 4. 3D AC view. Still working on this.

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 ($a \in \mathcal{A}$, $|\mathcal{A}| = p$) using a metric such as

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

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 and 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 $(i,j) \in \mathcal{N}$ and a fixed attribute $a \in \mathcal{A}$. $\mathcal{F}_X$ represents any data distribution with mean μ_X and variance σ_X^2.

It is clear that $|X_{ia} - X_{ja}|^q = |d_{ij}(a)|^q$ is another random variable. Let $Z_a^q \sim \mathcal{F}_{Z^q}(\mu_{z^q}, \sigma_{z^q}^2)$ be the random variable such that

$$Z_a^q = |d_{ij}(a)|^q = |X_{ia} - X_{ja}|^q, \quad a \in \mathcal{A}. \tag{4}$$

Furthermore, the collection $\{Z_a^q | a \in \mathcal{A}\}$ is a random sample of size p of mutually independent random variables. Hence, the sum of Z_a^q over all $a \in \mathcal{A}$ is asymptotically normal by the Classical Central Limit Theorem (CCLT). More explicitly, this implies that

$$\left(D_{ij}^{(q)}\right)^q = \sum_{a \in \mathcal{A}} |d_{ij}(a)|^q = \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q = \sum_{a \in \mathcal{A}} Z_a^q \dot{\sim} \mathcal{N}\left(\mu_{z^q} p, \sigma_{z^q}^2 p\right) \tag{5}$$

Consider the smooth function $g(z) = z^{1/q}$ that is continuously differentiable for $z > 0$. Assuming that $\mu_{z^q} > 0$, the Delta Method [3] can be applied to show that

$$\begin{aligned} g\left(\left(D_{ij}^{(q)}\right)^q\right) &= g\left(\sum_{a \in \mathcal{A}}^{p} Z_a^q\right) \\ &= \left(\sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q\right)^{1/q} \\ &= D_{ij}^{(q)} \dot{\sim} \mathcal{N}\left(g\left(\mu_{z^q} p\right), \left[g\prime\left(\mu_{z^q} p\right)\right]^2 \sigma_{z^q}^2 p\right) \\ \Rightarrow \quad D_{ij}^{(q)} &\dot{\sim} \mathcal{N}\left(\left(\mu_{z^q} p\right)^{1/q}, \frac{\sigma_{z^q}^2 p}{q^2 \left(\mu_{z^q} p\right)^{2\left(1-\frac{1}{q}\right)}}\right) \end{aligned} \tag{6}$$

Therefore, the distance between two fixed, distinct instances i and j given by Eq. 1 is asymptotically normal. Specifically, when $q = 2$, the distribution of $D_{ij}^{(2)}$ asymptotically approaches $\mathcal{N}\left(\sqrt{\mu_{z^2} p}, \frac{\sigma_{z^2}^2}{4\mu_{z^2}}\right)$. When p is small, however, we observe empirically that a closer estimate of the sample mean is

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2)}\right) &= \sqrt{\mathrm{E}\left[\left(D_{ij}^{(2)}\right)^2\right] - \mathrm{Var}\left(D_{ij}^{(2)}\right)} \\ &= \sqrt{\mu_{z^2} p - \frac{\sigma_{z^2}^2}{4\mu_{z^2}}}. \end{aligned} \tag{7}$$

One can readily verify the normality of distances between independent instances through sampling from any data distribution and plotting the histogram of pairwise distances. Histograms for the case of standard normal data and Manhattan $(q = 1)$ and

Euclidean ($q = 2$) metrics are shown in Figs. 5 and 6, respectively. For these simulated distances, we fixed the number of instances $m = 100$ and varied the number of attributes p from 10 to 10000. For even a rather small number of predictors, as in the case of $p = 10$, the sample distances are approximately normal. As p increases, the normality becomes stronger.

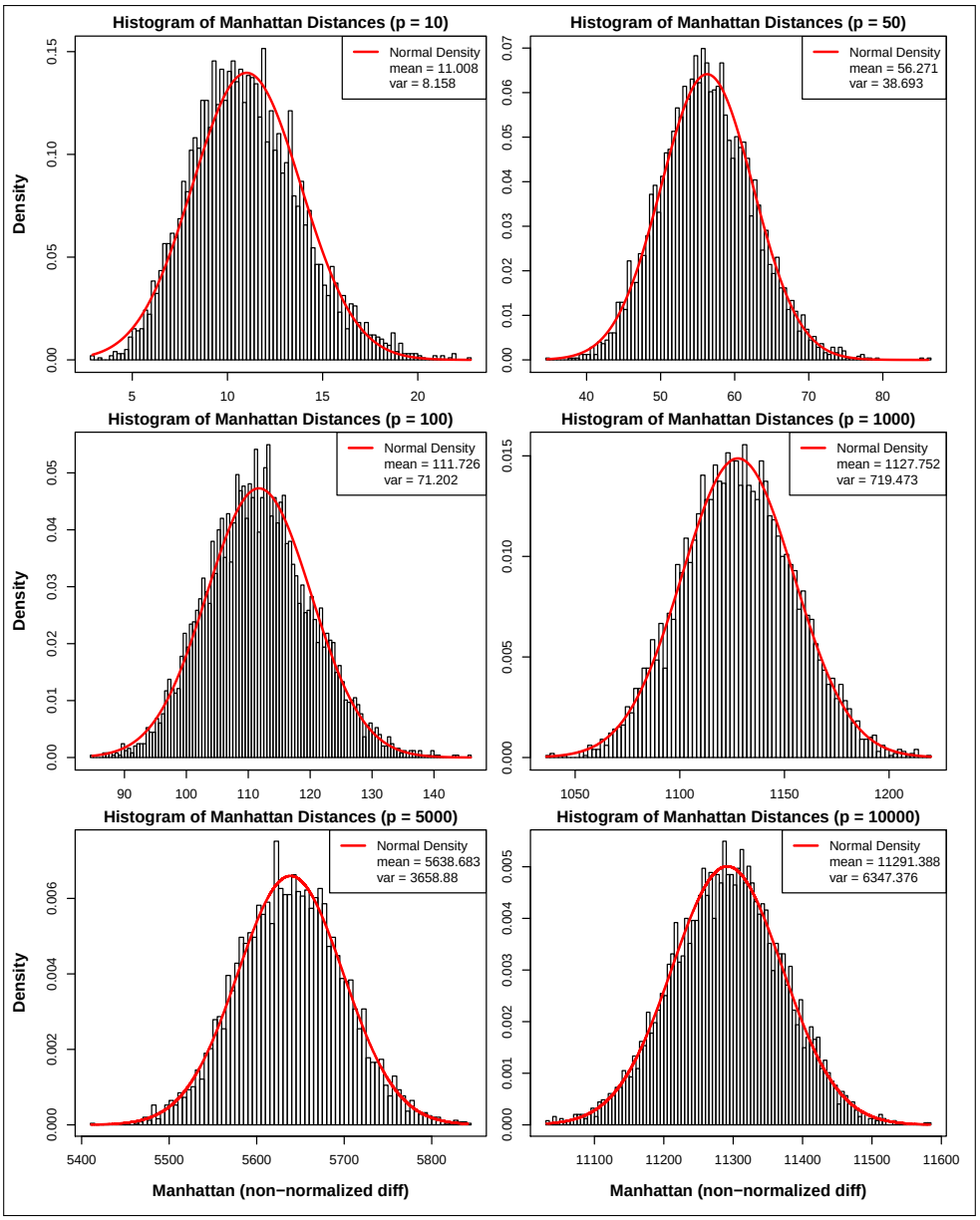

Fig 5. Convergence to normality of Manhattan distances between iid random normal instances. For each simulated distance distribution, we fixed $m = 100$ instances but varied p from 10 to 10000. It is clear that convergence is rapid, and approximate normality can be safely assumed for even $p = 10$.

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

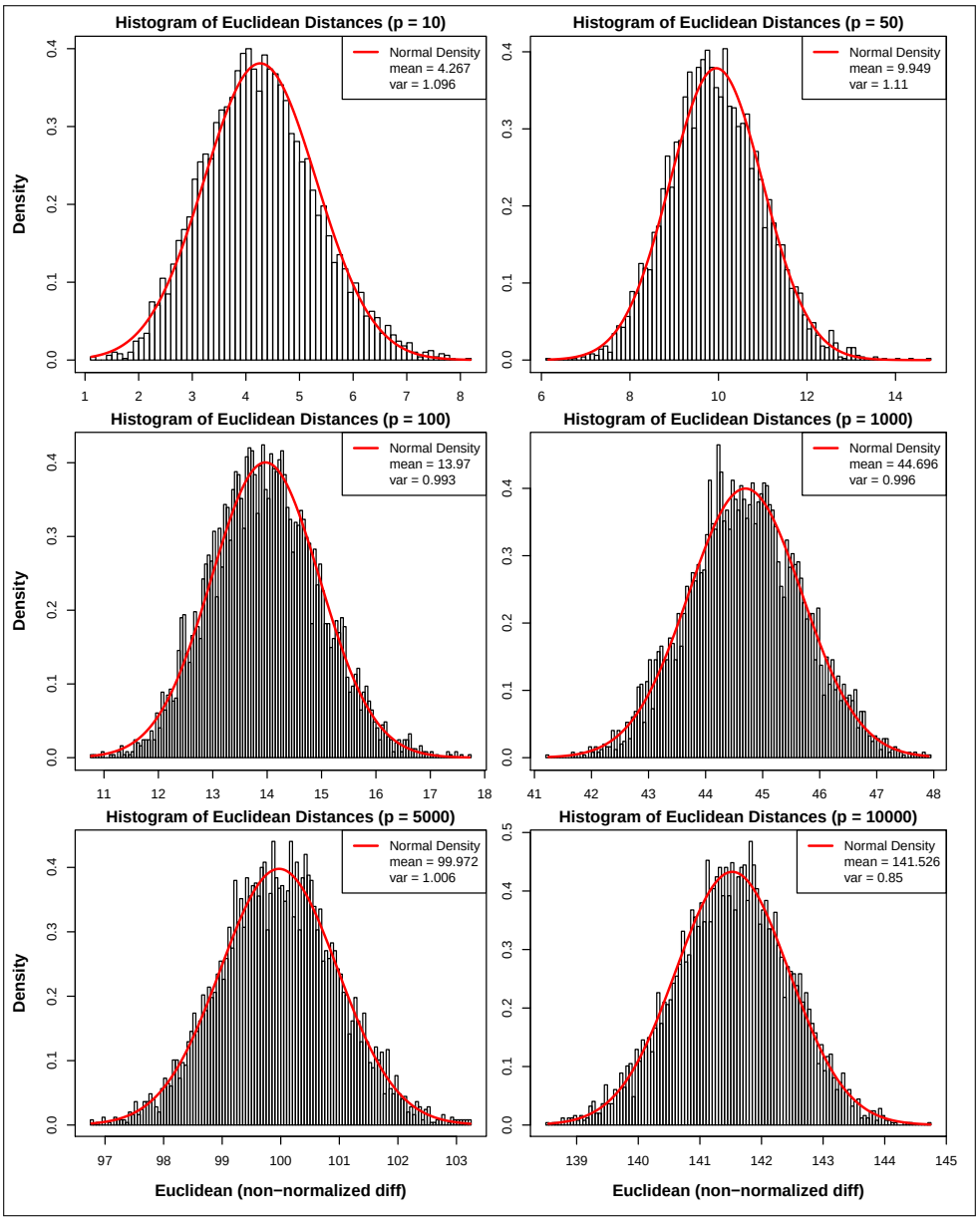

Fig 6. Convergence to normality of Euclidean distances between iid random normal instances. For each simulated distance distribution, we fixed $m = 100$ instances but varied p from 10 to 10000. It is clear that convergence is rapid, and approximate normality can be safely assumed for even $p = 10$.

distance matrix calculation. However, this collection is still asymptotically normal with mean and variance approximately equal to those given in Eq. 6. 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 some function of 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

$$D_{ij} \leq R_i^\alpha \implies j \in N_i^\alpha, \tag{8}$$

where the threshold radius for instance i is

$$R_i^\alpha = \bar{D}_i - \alpha\, \sigma_{\bar{D}_i} \tag{9}$$

and

$$\bar{D}_i = \frac{1}{m-1} \sum_{j \neq i} D_{ij}^{(\cdot)} \tag{10}$$

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

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

$$\bar{k}_\alpha = (m-1)q_\alpha, \tag{11}$$

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. An illustration of the area computed to estimate q_α is given by Fig. 7. This integral can be written in terms of the error function (erf):

$$q_\alpha = \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right). \tag{12}$$

And finally using Eqs. (11 and 12) we find

$$\bar{k}_\alpha = \left\lfloor \frac{m-1}{2}\left(1 - \mathrm{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right)\right\rfloor, \tag{13}$$

where we apply the floor function 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 $\bar{k}_\alpha^{\mathrm{hit/miss}} = \bar{k}_{1/2}^{\mathrm{hit/miss}} = \frac{1}{2}\bar{k}_{1/2} = .154(m-1)$,

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.

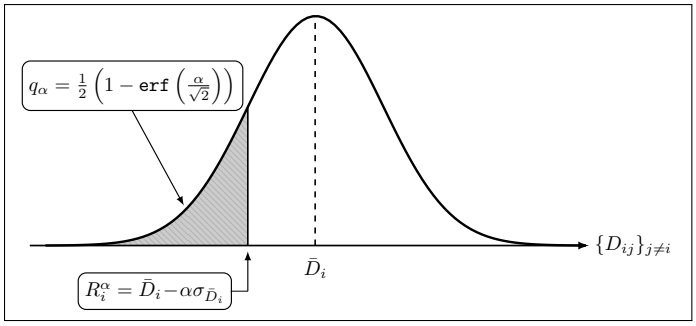

Fig 7. Illustration of the expected probability of a fixed instance j being in the fixed radius neighborhood of another instance i. The fixed radius is parameterized by a fraction α of the standard deviation of all pairwise distances measured from instance i to all possible neighbors.

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

3.1 Distribution of $|d_{ij}(a)|^q = |X_{ia} - X_{ja}|^q$

Suppose that $X_{ia}, X_{ja} \overset{iid}{\sim} \mathcal{F}_X(\mu_x, \sigma_x^2)$ and define $Z_a^q = |d_{ij}(a)|^q = |X_{ia} - X_{ja}|^q$, where $a \in \mathcal{A}$ and $|\mathcal{A}| = p$. In order to find the distribution of Z_a^q, we will use the following theorem given in [4].

Theorem 3.1 *Let $f(x)$ be the value of the probability density of the continuous random variable X at x. If the function given by $y = u(x)$ is differentiable and either increasing or decreasing for all values within the range of X for which $f(x) \neq 0$, then, for these values of x, the equation $y = u(x)$ can be uniquely solved for x to give $x = w(y)$, and for the corresponding values of y the probability density of $Y = u(X)$ is given by*

$$g(y) = f[w(y)] \cdot |w'(y)| \quad \textit{provided } u'(x) \neq 0$$

Elsewhere, $g(y) = 0$.

We have the following cases that result from solving for X_{ja} in the equation given by $Z_a^q = |X_{ia} - X_{ja}|^q$.

(i) Suppose that $X_{ja} = X_{ia} - (Z_a^q)^{1/q}$. Based on the iid assumption for X_{ia} and X_{ja}, it follows from Thm. 3.1 that the joint density function $g^{(1)}$ of X_{ia} and Z_a^q is given by

$$\begin{aligned} g^{(1)}(x_{ia}, z_a) &= f_X(x_{ia}, x_{ja}) \left| \frac{\partial x_{ja}}{\partial z_a} \right| \\ &= f_X(x_{ia}) f_X(x_{ja}) \left| \frac{-1}{q} \left(z_a^q\right)^{\frac{1}{q}-1} \right| \\ &= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}} f_X(x_{ia}) f_X\left(x_{ia} - \left(z_a^q\right)^{1/q}\right), \quad z_a > 0 \end{aligned} \tag{14}$$

The density function $f^{(1)}_{Z^q}$ of Z^q_a is then defined as

$$\begin{aligned} f^{(1)}_{Z^q}(z^q_a) &= \int_{-\infty}^{\infty} g^{(1)}(x_{ia}, z^q_a)\mathrm{d}x_{ia} \\ &= \frac{1}{q\left(z^q_a\right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) f_X\left(x_{ia} - \left(z^q_a\right)^{1/q}\right)\mathrm{d}x_{ia}, \quad z_a > 0 \end{aligned} \tag{15}$$

(ii) Suppose that $X_{ja} = X_{ia} + \left(Z^q_a\right)^{1/q}$. Based on the iid assumption for X_{ia} and X_{ja}, it follows from Thm. 3.1 that the joint density function $g^{(2)}$ of X_{ia} and Z_a is given by

$$\begin{aligned} g^{(2)}(x_{ia}, z_a) &= f_X(x_{ia}, x_{ja})\left|\frac{\partial x_{ja}}{\partial z_a}\right| \\ &= f_X(x_{ia}) f_X(x_{ja})\left|\frac{1}{q}\left(z^q_a\right)^{\frac{1}{q}-1}\right| \\ &= \frac{1}{q\left(z^q_a\right)^{1-\frac{1}{q}}} f_X(x_{ia}) f_X\left(x_{ia} - \left(z^q_a\right)^{1/q}\right), \quad z_a > 0 \end{aligned} \tag{16}$$

The density function $f^{(2)}_{Z^q}$ of Z^q_a is then defined as

$$\begin{aligned} f^{(2)}_{Z^q}(z^q_a) &= \int_{-\infty}^{\infty} g^{(2)}(x_{ia}, z^q_a)\mathrm{d}x_{ia} \\ &= \frac{1}{q\left(z^q_a\right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) f_X\left(x_{ia} + \left(z^q_a\right)^{1/q}\right)\mathrm{d}x_{ia}, \quad z_a > 0 \end{aligned} \tag{17}$$

Let F_{Z^q} denote the distribution function of the random variable Z^q_a. Furthermore, we define the events $E^{(1)}$ and $E^{(2)}$ as

$$E^{(1)} = \left\{|X_{ia} - X_{ja}|^q \le z^q_a | X_{ja} = X_{ia} - \left(Z^q_a\right)^{1/q}\right\} \tag{18}$$

and

$$E^{(2)} = \left\{|X_{ia} - X_{ja}|^q \le z^q_a | X_{ja} = X_{ia} + \left(Z^q_a\right)^{1/q}\right\}. \tag{19}$$

Then it follows from fundamental rules of probability that

$$\begin{aligned} F_{Z^q}(z^q_a) &= \mathrm{P}\left[Z^q_a \le z^q_a\right] \\ &= \mathrm{P}\left[|X_{ia} - X_{ja}|^q \le z^q_a\right] \\ &= \mathrm{P}\left[E^{(1)} \cup E^{(2)}\right] \\ &= \mathrm{P}\left[E^{(1)}\right] + \mathrm{P}\left[E^{(2)}\right] - \mathrm{P}\left[E^{(1)} \cap E^{(2)}\right] \\ &= \mathrm{P}\left[E^{(1)}\right] + \mathrm{P}\left[E^{(2)}\right] \\ &= \int_{-\infty}^{z^q_a} f^{(1)}_{Z^q}(t)\mathrm{d}t + \int_{-\infty}^{z^q_a} f^{(2)}_{Z^q}(t)\mathrm{d}t \\ &= \int_{-\infty}^{z^q_a} \left(f^{(1)}_{Z^q}(t) + f^{(2)}_{Z^q}(t)\right)\mathrm{d}t \\ &= \frac{1}{q\left(z^q_a\right)^{1-\frac{1}{q}}} \int_{-\infty}^{z^q_a} \left(\int_{-\infty}^{\infty} f_X(x_{ia})\left[f_X(x_{ia} - t) + f_X(x_{ia} + t)\right]\mathrm{d}x_{ia}\right)\mathrm{d}t, \quad z_a > 0 \end{aligned} \tag{20}$$

It follows directly from the result in Eq. 20 that the density function of the random variable Z_a^q is given by

$$\begin{aligned} f_{Z^q}(z_a^q) &= \frac{\partial}{\partial z_a^q} F_{Z^q}(z_a^q) \\ &= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[f_X\left(x_{ia} - \left(z_a^q\right)^{1/q}\right) + f_X\left(x_{ia} + \left(z_a^q\right)^{1/q}\right) \right] \mathrm{d}x_{ia} \end{aligned} \tag{21}$$

where $z_a > 0$.

Using Eq. 21, we can compute the mean and variance of the random variable Z_a^q as

$$\mu_{z^q} = \int_{-\infty}^{\infty} z_a^q f_{Z^q}(z_a^q) \mathrm{d}z_a^q \tag{22}$$

and

$$\sigma_{z^q}^2 = \int_{-\infty}^{\infty} \left(z_a^q\right)^2 f_{Z^q}(z_a^q) \mathrm{d}z_a^q - \mu_{z^q}^2. \tag{23}$$

It follows immediately from Eqs. 22 and 23 and the Classical Central Limit Theorem (CCLT) that

$$\left(D_{ij}^{(q)}\right)^q = \sum_{a \in \mathcal{A}} Z_a^q = \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \dot{\sim} \mathcal{N}\left(\mu_{z^q} p, \sigma_{z^q}^2 p\right) \tag{24}$$

Applying the result given in Eq. 6, the distribution of $D_{ij}^{(q)}$ is given by

$$D_{ij}^{(q)} \dot{\sim} \mathcal{N}\left(\left(\mu_{z^q} p\right)^{1/q}, \frac{\sigma_{z^q}^2 p}{q^2 \left(\mu_{z^q} p\right)^{2\left(1-\frac{1}{q}\right)}} \right), \quad \mu_{z^q} > 0 \tag{25}$$

with improved estimate of the mean for $q = 2$ given by Eq. 7.

3.1.1 Standard normal data

If $X_{ia}, X_{ja} \overset{iid}{\sim} \mathcal{N}(0,1)$, then the marginal density functions with respect to X for X_{ia}, $X_{ia} - \left(Z_a^q\right)^{1/q}$, and $X_{ia} + \left(Z_a^q\right)^{1/q}$ are defined as

$$f_X(x_{ia}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x_{ia}^2}, \tag{26}$$

$$f_X\left(x_{ia} - \left(z_a^q\right)^{1/q}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x_{ia} - \left(z_a^q\right)^{1/q}\right)^2}, \quad z_a > 0, \text{ and} \tag{27}$$

$$f_X\left(x_{ia} + \left(z_a^q\right)^{1/q}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x_{ia} + \left(z_a^q\right)^{1/q}\right)^2}, \quad z_a > 0 \tag{28}$$

Substituting the results given by Eqs. 26-28 into Eq. 21 and completing the square on x_{ia} in the exponents, we have

$$
\begin{aligned}
f_{Z^q}(z_a^q) &= \frac{1}{2q\pi\left(z_a^q\right)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \Bigg(e^{-\frac{1}{2}\left[\sqrt{2}x_{ia} - \frac{\sqrt{2}}{2}(z_a^q)^{1/q}\right]^2} \\
&\qquad\qquad + e^{-\frac{1}{2}\left[\sqrt{2}x_{ia} + \frac{\sqrt{2}}{2}(z_a^q)^{1/q}\right]^2} \Bigg) \mathrm{d}x_{ia} \qquad (29) \\
&= \frac{1}{2q\sqrt{\pi}\left(z_a^q\right)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left(e^{-\frac{1}{2}u^2} + e^{-\frac{1}{2}u^2} \right) \mathrm{d}u \qquad (30) \\
&= \frac{1}{2q\sqrt{\pi}\left(z_a^q\right)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} (1+1) \qquad (31) \\
&= \frac{1}{q\sqrt{\pi}} \left(z_a^q\right)^{\frac{1}{q}-1} e^{-\frac{1}{4}(z_a^q)^{2/q}} \qquad (32) \\
&= \frac{\frac{2}{q}}{(2^q)^{1/q}\,\Gamma\left(\frac{\frac{1}{q}}{\frac{2}{q}}\right)} \left(z_a^q\right)^{\frac{1}{q}-1} e^{-\left(\frac{z_a^q}{2^q}\right)^{2/q}} \qquad (33)
\end{aligned}
$$

The density function given by Eq. 29 is a Generalized Gamma density with parameters $b = \frac{2}{q}$, $c = 2^q$, and $d = \frac{1}{q}$. This distribution has mean and variance given by

$$
\begin{aligned}
\mu_{z^q} &= \frac{c\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \\
&= \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}
\end{aligned} \qquad (34)
$$

and

$$
\begin{aligned}
\sigma^2_{z^q} &= c^2 \left[\frac{\Gamma\left(\frac{d+2}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} - \left(\frac{\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \right)^2 \right] \\
&= 4^q \left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi} \right]
\end{aligned} \qquad (35)
$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 34 and 35 allow the p-dimensional mean and variance of the $D_{ij}^{(q)}$ distribution to be computed directly as

$$
\mu_{\left(D_{ij}^{(q)}\right)^q} = \mathrm{E}\left[\left(D_{ij}^{(q)}\right)^q\right] = \mathrm{E}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) = \sum_{a\in\mathcal{A}} \mathrm{E}\left(Z_a^q\right) = \sum_{a\in\mathcal{A}} \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} = \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} p \qquad (36)
$$

and

$$\begin{aligned}\sigma^2_{\left(D_{ij}^{(q)}\right)^q} = \operatorname{Var}\left[\left(D_{ij}^{(q)}\right)^q\right] &= \operatorname{Var}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) \\ &= \sum_{a\in\mathcal{A}} \operatorname{Var}\left(Z_a^q\right) \\ &= \sum_{a\in\mathcal{A}} 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right] \\ &= 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right] p\end{aligned} \tag{37}$$

Therefore, the asymptotic distribution of $D_{ij}^{(q)}$ for standard normal data is

$$\mathcal{N}\left(\left(2^q\frac{\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/q}, \frac{4^q p}{q^2\left(\frac{2^q\Gamma\left(\frac{1}{2}q+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]\right) \tag{38}$$

3.1.2 Standard uniform data

If $X_{ia}, X_{ja} \overset{iid}{\sim} \mathcal{U}(0,1)$, then the marginal density functions with respect to X for X_{ia}, $X_{ia} - \left(Z_a^q\right)^{1/q}$, and $X_{ia} + \left(Z_a^q\right)^{1/q}$ are defined as

$$f_X(x_{ia}) = 1, \quad 0 \le x_{ia} \le 1 \tag{39}$$

$$f_X\left(x_{ia} - \left(z_a^q\right)^{1/q}\right) = 1, \quad 0 \le x_{ia} - \left(z_a^q\right)^{1/q} \le 1, \text{ and} \tag{40}$$

$$f_X\left(x_{ia} + \left(z_a^q\right)^{1/q}\right) = 1, \quad 0 \le x_{ia} + \left(z_a^q\right)^{1/q} \le 1. \tag{41}$$

Substituting the results given by Eqs. 39-41 into Eq. 21, we have

$$\begin{aligned} f_{Z^q}(z_a^q) &= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}}\int_{-\infty}^{\infty} f_X(x_{ia})\left[f_X\left(x_{ia} - \left(z_a^q\right)^{1/q}\right) + f_X\left(x_{ia} + \left(z_a^q\right)^{1/q}\right)\right]\mathrm{d}x_{ia}, \\ &\qquad\qquad 0 < z_a \le 1 \\ &= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}}\int_0^1 [f_X(x_{ia} - \left(z_a^q\right) + f_X\left(x_{ia} + \left(z_a^q\right)^{1/q}\right)]\mathrm{d}x_{ia}, \quad 0 < z_a \le 1 \\ &= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}}\int_{\left(z_a^q\right)}^1 1\mathrm{d}x_{ia} + \int_0^{1-\left(z_a^q\right)} 1\mathrm{d}x_{ia}, \quad 0 < z_a \le 1 \\ &= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}}\left[\left(1-\left(z_a^q\right)\right) + \left(1-\left(z_a^q\right)\right)\right], \quad 0 < z_a \le 1 \\ &= \frac{1}{q}\cdot 2\left(z_a^q\right)^{\frac{1}{q}-1}\left[1-\left(z_a^q\right)^{1/q}\right]^{2-1}, \quad 0 < z_a \le 1 \end{aligned} \tag{42}$$

The density given by Eq. 42 is a Kumaraswamy density with parameters $b = \frac{1}{q}$ and $c = 2$ with moment generating function (MGF) given by

$$
\begin{aligned}
M_n &= \frac{c\Gamma\left(1+\frac{n}{b}\right)\Gamma(c)}{\Gamma\left(1+c+\frac{n}{b}\right)} \\
&= \frac{2}{(nq+2)(nq+1)}
\end{aligned}
\tag{43}
$$

Using the MGF given by Eq. 43, the mean and variance of Z_a^q are computed as

$$
\mu_{z^q} = \frac{2}{(q+2)(q+1)} \tag{44}
$$

and

$$
\sigma_{z^q}^2 = \frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2 \tag{45}
$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 46 and 47 allow the p-dimensional mean and variance of the $\left(D_{ij}^{(q)}\right)^q$ distribution to be computed directly as

$$
\begin{aligned}
\mu_{\left(D_{ij}^{(q)}\right)^q} = \mathrm{E}\left[\left(D_{ij}^{(q)}\right)^q\right] &= \mathrm{E}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) \\
&= \sum_{a\in\mathcal{A}} \mathrm{E}(Z_a^q) \\
&= \sum_{a\in\mathcal{A}} \frac{2}{(q+2)(q+1)} \\
&= \frac{2p}{(q+2)(q+1)}
\end{aligned}
\tag{46}
$$

and

$$
\begin{aligned}
\sigma^2_{\left(D_{ij}^{(q)}\right)^q} = \mathrm{Var}\left[\left(D_{ij}^{(q)}\right)^q\right] &= \mathrm{Var}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) \\
&= \sum_{a\in\mathcal{A}} \mathrm{Var}\left(Z_a^q\right) \\
&= \sum_{a\in\mathcal{A}} \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right] \\
&= \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right] p
\end{aligned}
\tag{47}
$$

Therefore, the asymptotic distribution of $D_{ij}^{(q)}$ for standard uniform data is

$$
\mathcal{N}\left(\left(\frac{2p}{(q+2)(q+1)}\right)^{1/q}, \right.
\left.
\frac{p}{q^2\left(\frac{2p}{(q+2)(q+1)}\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right]\right).
\tag{48}
$$

3.2 Manhattan ($q = 1$)

With our general formulas for the asymptotic mean and variance given by Eqs. 38 and 48 for any value of $q \in \mathbb{Z}^+$, we can simply substitute a particular value of q in order to determine the asymptotic distribution of the corresponding distance metric $D_{ij}^{(q)}$. We demonstrate this with the example of the Manhattan ($q = 1$) metric for standard normal and standard uniform data.

3.2.1 Standard normal data

Using the mean given by Eq. 38 and substituting $q = 1$, we have the following for standard normal data

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(1)}\right) &= \left(2\frac{\Gamma\left(\frac{1+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/1} \\ &= \frac{2p}{\sqrt{\pi}}\Gamma(1) \\ &= \frac{2p}{\sqrt{\pi}} \end{aligned} \tag{49}$$

Similarly, the variance of $D_{ij}^{(1)}$ is given by

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(1)}\right) &= \frac{4^1 p}{1^2\left(\frac{2^1\Gamma\left(\frac{1}{2}(1)+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{1}\right)}}\left[\frac{\Gamma\left(1+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}(1)+\frac{1}{2}\right)}{\pi}\right] \\ &= \frac{4p}{1}\left[\frac{\frac{1}{2}\Gamma\left(\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2(1)}{\pi}\right] \\ &= 4p\left[\frac{1}{2} - \frac{1}{\pi}\right] \\ &= \frac{2(\pi-2)p}{\pi} \end{aligned} \tag{50}$$

3.2.2 Standard uniform data

Using the mean given by Eq. 48 and substituting $q = 1$, we have the following for standard uniform data

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(1)}\right) &= \left(\frac{2p}{(1+2)(1+1)}\right)^{1/1} \\ &= \frac{2p}{6} \\ &= \frac{p}{3} \end{aligned} \tag{51}$$

Similarly, the variance of $D_{ij}^{(1)}$ is given by

$$\begin{aligned}
\operatorname{Var}\left(D_{ij}^{(1)}\right) &= \frac{p}{1^2\left(\frac{2p}{(1+2)(1+1)}\right)^{2\left(1-\frac{1}{1}\right)}}\left[\frac{1}{(1+1)(2(1)+1)} - \left(\frac{2}{(1+2)(1+1)}\right)^2\right] \\
&= p\left[\frac{1}{6} - \frac{1}{9}\right] \\
&= \frac{p}{18}
\end{aligned} \tag{52}$$

3.3 Euclidean ($q = 2$)

Analogous to the previous section, we demonstrate the usage of Eqs. 38 and 48 for the Euclidean ($q = 2$) metric for standard normal and standard uniform data.

3.3.1 Standard normal data

Using the mean given by Eq. 38 and substituting $q = 2$, we have the following for standard normal data

$$\begin{aligned}
\operatorname{E}\left(D_{ij}^{(2)}\right) &= \left(2\frac{\Gamma\left(\frac{2+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/2} \\
&= \left(\frac{2p}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right)\right)^{1/2} \\
&= \sqrt{2p}
\end{aligned} \tag{53}$$

Similarly, the variance of $D_{ij}^{(2)}$ is given by

$$\begin{aligned}
\operatorname{Var}\left(D_{ij}^{(1)}\right) &= \frac{4^2 p}{2^2\left(\frac{2^2\Gamma\left(\frac{1}{2}(2)+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{2}\right)}}\left[\frac{\Gamma\left(2+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}(2)+\frac{1}{2}\right)}{\pi}\right] \\
&= \frac{16p}{4\left(\frac{4\Gamma\left(\frac{3}{2}\right)}{\sqrt{\pi}}p\right)}\left[\frac{\Gamma\left(\frac{5}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{3}{2}\right)}{\pi}\right] \\
&= 2\left[\frac{3}{4} - \frac{1}{4}\right] \\
&= 1
\end{aligned} \tag{54}$$

For the case in which the number of attributes p is small, an improved estimate of the mean is given by Eq. 7. The lower dimensional estimate of the mean is as follows.

$$\begin{aligned}
\operatorname{E}\left(D_{ij}^{(2)}\right) &= \left(2\frac{\Gamma\left(\frac{2+1}{2}\right)}{\sqrt{\pi}}p - 1\right)^{1/2} \\
&= \left(\frac{2p}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) - 1\right)^{1/2} \\
&= \sqrt{2p-1}
\end{aligned} \tag{55}$$

For high dimensional data sets, such as gene expression, rs-fMRI, or GWAS, it is clear that the magntiude of p will be sufficient to use Eq. 53 since $\sqrt{2p} \approx \sqrt{2p-1}$ in that case.

3.3.2 Standard uniform data

Using the mean given by Eq. 48 and substituting $q = 2$, we have the following for standard uniform data

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2)}\right) &= \left(\frac{2p}{(2+2)(2+1)}\right)^{1/2} \\ &= \left(\frac{2p}{12}\right)^{1/2} \\ &= \sqrt{\frac{p}{6}} \end{aligned} \tag{56}$$

Similarly, the variance of $D_{ij}^{(2)}$ is given by

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(2)}\right) &= \frac{p}{2^2\left(\frac{2p}{(2+2)(2+1)}\right)^{2\left(1-\frac{1}{2}\right)}}\left[\frac{1}{(2+1)(2(2)+1)} - \left(\frac{2}{(2+2)(2+1)}\right)^2\right] \\ &= \frac{3p}{2}\left[\frac{1}{15} - \frac{1}{36}\right] \\ &= \frac{7p}{120} \end{aligned} \tag{57}$$

For the case in which the number of attributes p is small, an improved estimate of the mean is given by Eq. 7. The lower dimensional estimate of the mean is as follows.

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2)}\right) &= \left(\frac{2p}{(2+2)(2+1)} - \frac{7}{120}\right)^{1/2} \\ &= \left(\frac{2p}{12} - \frac{7}{120}\right)^{1/2} \\ &= \sqrt{\frac{p}{6} - \frac{7}{120}} \end{aligned} \tag{58}$$

For high dimensional data sets, such as gene expression, rs-fMRI, or GWAS, it is clear that the magntiude of p will be sufficient to use Eq. 53 since $\sqrt{\frac{p}{6}} \approx \sqrt{\frac{p}{6} - \frac{7}{120}}$ in that case.

3.4 Distribution of attribute extremes

For Relief-based methods [5,6], the standard numeric diff metric is given by

$$d_{ij}^{\mathrm{num}}(a) = \mathrm{diff}(a,(i,j)) = \frac{|X_{ia} - X_{ja}|}{\max(a) - \min(a)} \tag{59}$$

where $\max(a) = \max_{k\in\mathcal{I}}\{X_{ka}\}$, $\min(a) = \min_{k\in\mathcal{I}}\{X_{ka}\}$, and $\mathcal{I} = \{1,2,\ldots,m\}$.

In order to determine moments of asymptotic distance distributions induced by Eq. 59, we must first derive the asymptotic extreme value distributions of the attribute maximum and minimum. Although the exact distribution of the maximum or minimum requires an assumption about the data distribution, the Fisher-Tippett-Gnedenko Theorem allows us to categorize the extreme value distribution for a collection of independent and identically distributed random variables into one of three distributional families. Before stating the theorem, we first need the following definition.

Definition 3.1 *A distribution $\mathcal{F}_X$ is said to be **degenerate** if its density function f_X is the Dirac delta $\delta(x - c_0)$ centered at a constant $c_0 \in \mathbb{R}$, with corresponding distribution function F_X defined as*

$$F_X(x) = \begin{cases} 1, & x \geq c_0, \\ 0, & x < c_0. \end{cases}$$

Theorem 3.2 (Fisher-Tippett-Gnedenko) *Let $X_{1a}, X_{2a}, \ldots, X_{ma} \overset{iid}{\sim} \mathcal{F}_X\left(\mu_x, \sigma_x^2\right)$ and let $X_a^\alpha = \max_{k \in \mathcal{I}}\{X_{ka}\}$. If there exists two non-random sequences $b_m > 0$ and c_m such that*

$$\lim_{m \to \infty} P\left(\frac{X_a^\alpha - c_m}{b_m} \leq x\right) = G_X(x),$$

where G_X is a non-degenerate distribution function, then the limiting distribution $\mathcal{G}_X$ is in the Gumbel, Fréchet, or Wiebull family.

The three distribution families given in Thm. 3.2 are actually special cases of the Generalized Extreme Value Distribution. In the context of extreme values, Thm. 3.2 is analogous to the Central Limit Theorem for the distribution of sample mean. We will take advantage of this theorem for the distribution of the maximum for standard normal data to show that the limiting distribution is in the Gumbel family. However, we will derive the distribution of the maximum and minimum for standard uniform data directly. Regardless of data type, the distribution of the sample maximum is derived in Eq. 60.

$$\begin{aligned} \mathrm{P}[X_a^\alpha \leq x] &= \mathrm{P}\left[\max_{k \in \mathcal{I}}\{X_{ka}\} \leq x\right] \\ &= \mathrm{P}[X_{1a} \leq x, X_{2a} \leq x, \ldots, X_{ma} \leq x] \\ &= \prod_{k=1}^{m} \mathrm{P}[X_{ka} \leq x] \\ &= \prod_{k=1} F_X(x) \\ &= [F_X(x)]^m \end{aligned} \tag{60}$$

Therefore, we have the following expression for the distribution function of the maximum.

$$F_{\max}(x) = [F_X(x)]^m \tag{61}$$

Differentiating the distribution function given by Eq. 61 gives us the following density function for the distribution of the maximum.

$$\begin{aligned} f_{\max}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{\max}(x) \\ &= \frac{\mathrm{d}}{\mathrm{d}x} [F_X(x)]^m \\ &= m[F_X(x)]^{m-1} f_X(x) \end{aligned} \tag{62}$$

The distribution of the sample minimum, X_a^ω, is derived in Eq. 63.

$$
\begin{aligned}
\mathrm{P}[X_a^\omega \leq x] &= 1 - \mathrm{P}[X_a^\omega \geq x] \\
&= 1 - \mathrm{P}\left[\min_{k\in\mathcal{I}}\{X_{ka}\} \geq x\right] \\
&= 1 - \mathrm{P}[X_{1a} \geq x, X_{2a} \geq x, \ldots, X_{ma} \geq x] \\
&= 1 - \prod_{k=1}^{m} \mathrm{P}[X_{ka} \geq x] \\
&= 1 - [\mathrm{P}[X_{1a} \geq x]]^m \\
&= 1 - [1 - \mathrm{P}[X_{1a} \leq x]]^m \\
&= 1 - [1 - F_X(x)]^m
\end{aligned}
\tag{63}
$$

Therefore, we have the following expression for the distribution function of the minimum.

$$
F_{\min}(x) = 1 - [1 - F_X(x)]^m \tag{64}
$$

Differentiating the distribution function given by Eq. 64 gives us the following density function for the distribution of the minimum.

$$
\begin{aligned}
f_{\min}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{\min}(x) \\
&= \frac{\mathrm{d}}{\mathrm{d}x} \left(1 - [1 - F_X(x)]^m\right) \\
&= m\,[1 - F_X(x)]^{m-1}\, f_X(x)
\end{aligned}
\tag{65}
$$

Given the densities of the distribution of sample maximum and minimum, we can easily compute moments and the variance. The first and second moment about the origin and the variance of the distribution of the maximum are given by the following.

$$
\begin{aligned}
\mu_\alpha^{(1)}(m) = \mathrm{E}(X_a^\alpha) &= \int_{-\infty}^{\infty} x f_{\max}(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x \left(m[F_X(x)]^{m-1} f_X(x)\right) \mathrm{d}x \\
&= m \int_{-\infty}^{\infty} x f_X(x)[F_X(x)]^{m-1}\mathrm{d}x
\end{aligned}
\tag{66}
$$

$$
\begin{aligned}
\mu_\alpha^{(2)}(m) = \mathrm{E}[(X_a^\alpha)^2] &= \int_{-\infty}^{\infty} x^2 f_{\max}(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x^2 \left(m[F_X(x)]^{m-1} f_X(x)\right) \mathrm{d}x \\
&= m \int_{-\infty}^{\infty} x^2 f_X(x)[F_X(x)]^{m-1}\mathrm{d}x
\end{aligned}
\tag{67}
$$

$$
\sigma_\alpha^2(m) = \mu_\alpha^{(2)}(m) - \left[\mu_\alpha^{(1)}(m)\right]^2 \tag{68}
$$

Similarly, we have the first and second moment about the origin and variance of the distribution of sample minimum given by the following.

$$\begin{aligned}\mu_{\omega}^{(1)}(m) = \mathrm{E}(X_a^{\omega}) &= \int_{-\infty}^{\infty} x f_{\min}(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x\left(m[F_X(x)]^{m-1} f_X(x)\right)\mathrm{d}x \\ &= m\int_{-\infty}^{\infty} x f_X(x)[F_X(x)]^{m-1}\mathrm{d}x\end{aligned} \tag{69}$$

$$\begin{aligned}\mu_{\omega}^{(2)}(m) = \mathrm{E}[(X_a^{\omega})^2] &= \int_{-\infty}^{\infty} x^2 f_{\min}(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2\left(m[F_X(x)]^{m-1} f_X(x)\right)\mathrm{d}x \\ &= m\int_{-\infty}^{\infty} x^2 f_X(x)[F_X(x)]^{m-1}\mathrm{d}x\end{aligned} \tag{70}$$

$$\sigma_{\omega}^2(m) = \mu_{\omega}^{(2)}(m) - \left[\mu_{\omega}^{(1)}(m)\right]^2 \tag{71}$$

With the densities of attribute maximum and minimum for sample size m, the expected range is given by the following.

$$\begin{aligned}\mathrm{E}(X_a^{\alpha} - X_a^{\omega}) &= \mathrm{E}(X_a^{\alpha}) - \mathrm{E}(X_a^{\omega}) \\ &= \mu_{\alpha}^{(1)}(m) - \mu_{\omega}^{(1)}(m)\end{aligned} \tag{72}$$

For a data distribution that has zero skewness and has support that is symmetric about 0, the result given by Eq. 72 can be simplified to the following expression.

$$\mathrm{E}(X_a^{\alpha} - X_a^{\omega}) = 2\mu_{\alpha}^{(1)}(m) \tag{73}$$

For large samples ($m >> 1$), the covariance between the sample maximum and minimum is approximately zero [7]. Therefore, the variance of the attribute range of a sample of size m is given by the following.

$$\begin{aligned}\mathrm{Var}(X_a^{\alpha} - X_a^{\omega}) &\approx \mathrm{Var}(X_a^{\alpha}) + \mathrm{Var}(X_a^{\omega}) \\ &= \sigma_{\alpha}^2(m) + \sigma_{\omega}^2(m)\end{aligned} \tag{74}$$

Under the assumption of zero skewness and support that is symmetric about 0, the result given by Eq. 74 becomes the following.

$$\begin{aligned}\mathrm{Var}(X_a^{\alpha} - X_a^{\omega}) &= 2\mathrm{Var}(X_a^{\alpha}) \\ &= 2\sigma_{\alpha}^2\end{aligned} \tag{75}$$

Let $\mu_{D_{ij}^{(q)}}$ and $\sigma^2_{D_{ij}^{(q)}}$ denote the mean and variance given by Eq. 25. Furthermore, let $D_{ij}^{(q*)}$ denote the max-min normalized distance between instances i and j that is induced by the metric given by Eq. 59. Then the mean of the max-min normalized distance distribution is given by the following.

$$
\begin{aligned}
\mu_{D_{ij}^{(q*)}} &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia}-X_{ja}|}{X_a^\alpha - X_a^\omega}\right)^q\right)^{1/q}\right] \\
&\approx \frac{1}{\mathrm{E}(X_a^\alpha - X_a^\omega)}\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}|X_{ia}-X_{ja}|^q\right)^{1/q}\right] \\
&= \frac{\mu_{D_{ij}^{(q)}}}{\mathrm{E}(X_a^\alpha) - \mathrm{E}(X_a^\omega)} \\
&= \frac{\mu_{D_{ij}^{(q)}}}{\mu_\alpha^{(1)} - \mu_\omega^{(1)}}
\end{aligned}
\tag{76}
$$

The variance of the max-min normalized distance distribution is given by the following.

$$
\begin{aligned}
\sigma^2_{D_{ij}^{(q*)}} &= \mathrm{Var}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia}-X_{ja}|}{X_a^\alpha - X_a^\omega}\right)^q\right)^{1/q}\right] \\
&= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia}-X_{ja}|}{X_a^\alpha - X_a^\omega}\right)^q\right)^{2/q}\right] - \left(\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia}-X_{ja}|}{X_a^\alpha - X_a^\omega}\right)^q\right)^{1/q}\right]\right)^2 \\
&\approx \frac{\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}|X_{ia}-X_{ja}|^q\right)^{2/q}\right]}{\mathrm{E}[(X_a^\alpha - X_a^\omega)^2]} - \frac{\left(\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}|X_{ia}-X_{ja}|^q\right)^{1/q}\right]\right)^2}{\mathrm{E}[(X_a^\alpha - X_a^\omega)^2]} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}} + \mu^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^\alpha - X_a^\omega)^2]} - \frac{\mu^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^\alpha - X_a^\omega)^2]} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^\alpha - X_a^\omega)^2]} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^\alpha)^2] - 2\mathrm{E}(X_a^\alpha)\mathrm{E}(X_a^\omega) + \mathrm{E}(X_a^\omega)} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}}}{\mu_\alpha^{(2)}(m) - 2\mu_\alpha^{(1)}(m)\mu_\omega^{(1)}(m) + \mu_\omega^{(2)}(m)}
\end{aligned}
\tag{77}
$$

With the results given by Eqs. 76 and 77, we have the following generalized estimate for the asymptotic distribution of the max-min normalized distance distribution.

$$
D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{\mu_\alpha^{(1)}(m) - \mu_\omega^{(1)}(m)}, \frac{\sigma^2_{D_{ij}^{(q)}}}{\mu_\alpha^{(2)}(m) - 2\mu_\alpha^{(1)}(m)\mu_\omega^{(1)}(m) + \mu_\omega^{(2)}(m)}\right) \tag{78}
$$

For data with zero skewness and support that is symmetric about 0, the expected sample maximum is the additive inverse of the expected sample minimum. This allows us to express the formula given by Eq. 76 exclusively in terms of the expected maximum. This result is given by the following.

$$
\mu_{D_{ij}^{(q*)}} \approx \frac{\mu_{D_{ij}^{(q)}}}{2\mu_\alpha^{(1)}(m)} \tag{79}
$$

A similar substitution gives us the following expression for the variance of the max-min normalized distance distribution.

$$\begin{aligned}\sigma^2_{D_{ij}^{(q*)}} &\approx \frac{\sigma^2_{D_{ij}^{(q)}}}{2\mu_\alpha^{(2)}(m) + 2\left[\mu_\alpha^{(1)}(m)\right]^2} \\ &= \frac{\sigma^2_{D_{ij}^{(q)}}}{2\left(\sigma_\alpha^2(m) + \left[\mu_\alpha^{(1)}(m)\right]^2\right)}\end{aligned} \tag{80}$$

Therefore, the asymptotic distribution of the max-min normalized distance distribution is given by the following.

$$D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu_\alpha^{(1)}(m)}, \frac{\sigma^2_{D_{ij}^{(q)}}}{2\left(\sigma_\alpha^2(m) + \left[\mu_\alpha^{(1)}(m)\right]^2\right)}\right) \tag{81}$$

3.4.1 Standard Normal Data

Standard normal data has zero skewness and has support that is symmetric about 0. This implies that the mean and variance of the distribution of sample range can be expressed exclusively in terms of the sample maximum. Given the nature of the density function of the sample maximum for sample size m, the integration required to determine the moments given by Eqs. 66 and 67 is not possible. These moments can either be approximated numerically or we can use extreme value theory to determine the form of the asymptotic distribution of the sample maximum. Using the latter method, we will show that the asymptotic distribution of the sample maximum for standard normal data is in the Gumbel family. Let $c_m = -\Phi^{-1}\left(\frac{1}{m}\right)$ and $b_m = \frac{1}{c_m}$. Using Taylor's Theorem, we have the following expansion.

$$\begin{aligned}\log\Phi(-c_m - b_m x) &= \log\Phi(-c_m) - b_m x\frac{\phi(-c_m)}{\Phi(-c_m)} + \mathcal{O}(b_m^2 x^2) \\ &= \log\left(\frac{1}{m}\right) - x\frac{\phi(-c_m)}{c_m\Phi(-c_m)} + \mathcal{O}(b_m^2 x^2)\end{aligned} \tag{82}$$

In order to simplify the right-hand side of Eq. 82, we will use the well known Mills Ratio Bounds [8] given by the following.

$$1 \leq \frac{\phi(x)}{x\Phi(-x)} \leq 1 + \frac{1}{x^2} \quad , x > 0 \tag{83}$$

The inequalities given by Eq. 83 show that $\frac{\phi(x)}{x\Phi(-x)} \to 1$ as $x \to \infty$. This implies that $\frac{\phi(c_m)}{c_m\Phi(-c_m)} \to 1$ as $m \to \infty$ since $c_m = -\Phi^{-1}\left(\frac{1}{m}\right) \to \infty$ as $m \to \infty$. This gives us the following approximation of the right-hand side of Eq. 82.

$$\begin{aligned}&\log\Phi(-c_m - b_m x) \approx \log\left(\frac{1}{m}\right) - x + \mathcal{O}(b_m^2 x^2) \\ &\Rightarrow \Phi(-c_m - b_m x) \approx \frac{1}{m}e^{-x+\mathcal{O}(b_m^2 x^2)} \\ &\Rightarrow \Phi(c_m + b_m x) \approx 1 - \frac{1}{m}e^{-x+\mathcal{O}(b_m^2 x^2)}\end{aligned} \tag{84}$$

Using the result given by Eq. 84, we have the following.

$$
\begin{aligned}
\mathrm{P}\left(\frac{X_a^\alpha - c_m}{b_m} \leq x\right) &= \mathrm{P}(X_a^\alpha \leq c_m + b_m x) \\
&= \Phi^m(c_m + b_m x) \\
&\approx \left(1 - \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)}\right)^m \\
&= \left(1 - \frac{1}{m} e^{-x + \mathcal{O}\left(\frac{1}{c_m^2} x^2\right)}\right)^m \\
&\approx \left(1 - \frac{1}{m} e^{-x}\right)^m \\
\Rightarrow \lim_{m\to\infty} \mathrm{P}\left(\frac{X_a^\alpha - c_m}{b_m} \leq x\right) &= \lim_{m\to\infty} \left(1 - \frac{1}{m} e^{-x}\right)^m \\
&= e^{-e^{-x}}
\end{aligned}
\tag{85}
$$

The right-hand side of Eq. 85 is the cumulative distribution function of the standard Gumbel distribution. The mean of the asymptotic distribution is given by the following.

$$
\mathrm{E}(X_a^\alpha) = \mu_\alpha^{(1)} = -\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)} \tag{86}
$$

where γ is the Euler-Mascheroni constant. The median of this distribution is given by the following.

$$
\widetilde{\mu}_\alpha = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right) \tag{87}
$$

Finally, the variance of the asymptotic distribution of the sample maximum is given by the following.

$$
\mathrm{Var}(X_a^\alpha) = \frac{\pi^2}{6}\left(\frac{1}{-\Phi^{-1}\left(\frac{1}{m}\right)}\right)^2 \tag{88}
$$

For typical sample sizes m in high-dimensional spaces, the variance estimate given by Eq. 88 exceeds the variance of the sample maximum significantly. Using the fact that $-\Phi^{-1}\left(\frac{1}{m}\right) \dot{\sim} \sqrt{2\log(m)}$ [9] and $\frac{1}{2\log(m)} \leq \left(\frac{1}{-\Phi^{-1}\left(\frac{1}{m}\right)}\right)^2$ for $m \geq 2$, we can get a more accurate approximation of the variance with the following.

$$
\begin{aligned}
\sigma_\alpha^2(m) = \mathrm{Var}(X_a^\alpha) &\approx \frac{\pi^2}{6}\left(\frac{1}{\sqrt{2\log(m)}}\right)^2 \\
&= \frac{\pi^2}{12\log(m)}
\end{aligned}
\tag{89}
$$

Then the mean of the range of m iid standard normal random variables are given by the following.

$$
\mathrm{E}(X_a^\alpha - X_a^\omega) = 2\mu_\alpha^{(1)}(m) = 2\left[-\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)}\right] \tag{90}
$$

It is well known that the sample extremes from the standard normal distribution are approximately uncorrelated for large sample size m [7]. This implies that we can

approximate the variance of the range of m iid standard normal random variables with the following result.

$$\begin{aligned}\mathrm{Var}(X_a^\alpha - X_a^\omega) &\approx \mathrm{Var}(X_a^\alpha) + \mathrm{Var}(X_a^\omega) \\ &= \sigma_\alpha^2(m) + \sigma_\omega^2(m) \\ &= 2\sigma_\alpha^2(m) \\ &\approx 2\left(\frac{\pi^2}{2\log(m)}\right) \\ &= \frac{\pi^2}{6\log(m)}\end{aligned} \tag{91}$$

For the purpose of approximating the mean and variance of the max-min normalized distance distribution, the formula for the median of the distribution of the attribute maximum yields more accurate results. That is, the approximation of the expected maximum given by Eq. 86 overestimates the sample maximum. The formula for the median of the sample maximum, given by Eq. 87, provides a more accurate estimate of this sample extreme. Therefore, the following estimate for the mean of the attribute range will be used instead.

$$\mathrm{E}(X_a^\alpha - X_a^\omega) = 2\mu_\alpha^{(1)}(m) \approx 2\left[\frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)\right] \tag{92}$$

We have already determined that $\mu_{D_{ij}^{(q)}}$ and $\sigma^2_{D_{ij}^{(q)}}$ are given by Eq. 38. Using the results given by Eqs. 92 and 91 and the general formulas for the mean and variance of the max-min normalized distance distribution given in Eq. 81, this leads us to the following asymptotic estimate for the distribution of the max-min normalized distances for standard normal data.

$$D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu_\alpha^{(1)}(m)}, \frac{6\log(m)\sigma^2_{D_{ij}^{(q)}}}{\pi^2 + 24\left[\mu_\alpha^{(1)}(m)\right]^2\log(m)}\right) \tag{93}$$

3.4.2 Standard Uniform Data

Standard uniform data does not have support that is symmetric about 0. Due to the simplicity of the density function, however, we can derive the distribution of the maximum and minimum of a sample of size m explicitly. Using the general forms of the distribution functions of the maximum and minimum given by Eqs. 61 and 64, we have the following distribution functions for standard uniform data.

$$F_{\max}(x) = x^m \tag{94}$$

$$F_{\min}(x) = 1 - (1-x)^m \tag{95}$$

Using the general forms of the density functions of the maximum and minimum given by Eqs. 62 and 65, we have the following density functions for standard uniform data.

$$f_{\max}(x) = mx^{m-1} \tag{96}$$

$$f_{\min}(x) = m(1-x)^{m-1} \tag{97}$$

Then the expected maximum and minimum are computed through straightforward integration as follows.

$$\begin{aligned} \mathrm{E}(X_a^\alpha) = \mu_\alpha^{(1)}(m) &= \int_0^1 x f_{\max}(x)\mathrm{d}x \\ &= \int_0^1 x[mx^{m-1}]\mathrm{d}x \\ &= \frac{m}{m+1} \end{aligned} \tag{98}$$

$$\begin{aligned} \mathrm{E}(X_a^\omega) = \mu_\omega^{(1)}(m) &= \int_0^1 x f_{\min}(x)\mathrm{d}x \\ &= \int_0^1 x[m(1-x)^{m-1}]\mathrm{d}x \\ &= \frac{1}{m+1} \end{aligned} \tag{99}$$

We can compute the second moment about the origin of the sample range as follows.

$$\begin{aligned} \mathrm{E}[(X_a^\alpha - X_a^\omega)^2] &= \mathrm{E}[(X_a^\alpha)^2 - 2X_a^\alpha X_a^\omega + (X_a^\omega)^2] \\ &= \mathrm{E}[(X_a^\alpha)^2] - 2\mathrm{E}(X_a^\alpha)\mathrm{E}(X_a^\omega) + \mathrm{E}[(X_a^\omega)^2] \\ &= \mu_\alpha^{(2)}(m) - 2\mu_\alpha^{(1)}(m)\mu_\omega^{(1)}(m) + \mu_\omega^{(2)}(m) \\ &= \int_0^1 x^2[mx^{m-1}]\mathrm{d}x - 2\left(\frac{m}{m+1}\right)\left(\frac{1}{m+1}\right) + \int_0^1 x^2[m(1-x)^{m-1}]\mathrm{d}x \\ &= \frac{m}{m+2} - \frac{2m}{(m+1)^2} + \frac{2}{(m+1)(m+2)} \\ &= \frac{m^3 - m + 2}{(m+2)(m+1)^2} \end{aligned} \tag{100}$$

Using the general formulas given in Eq. 78 and the mean ($\mu_{D_{ij}^{(q)}}$) and variance ($\sigma^2_{D_{ij}^{(q)}}$) given by Eq. 48, we have the following asymptotic estimate for the max-min normalized distance distribution for standard uniform data.

$$D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{(m+1)\mu_{D_{ij}^{(q)}}}{m-1}, \frac{(m+2)(m+1)^2\sigma^2_{D_{ij}^{(q)}}}{m^3-m+2}\right) \tag{101}$$

3.5 GWAS Distance Distributions

Consider a GWAS data set, which has the following encoding based on minor allele frequency.

$$X_{ia} = \begin{cases} 0 & \text{if there are no minor alleles at locus } a \\ 1 & \text{if there is 1 minor allele at locus } a \\ 2 & \text{if there are 2 minor alleles at locus } a \end{cases} \tag{102}$$

A minor allele at a particular locus a is the least frequent of the two alleles at that particular locus a. For random GWAS data sets, we can think X_{ia} as the number of successes in two Bernoulli trials. That is, $X_{ia} \sim \mathcal{B}(2, f_a)$ where f_a is the probability of success. The success probability f_a is the probability of a minor allele occurring at X_{ia}. Furthermore, the minor allele probabilities are assumed to be independent and

identically distributed. Two commonly known types of metrics for GWAS data are the Genotype Mismatch (GM) and Allele Mismatch (AM) metrics. The GM and AM metrics are defined as follows.

$$d_{ij}^{\mathrm{GM}}(a) = \begin{cases} 0 & \text{if } X_{ia} \neq X_{ja} \\ 1 & \text{otherwise} \end{cases} \tag{103}$$

$$d_{ij}^{\mathrm{AM}}(a) = \frac{1}{2}\big|X_{ia} - X_{ja}\big| \tag{104}$$

A more informative metric must take into account whether differences in allele frequency at a particular locus a result in transitions or transversions. A metric that accounts for transitions (Ti) and transversions (Tv) was introduced in [10]. This metric is given by the following.

$$d_{ij}^{\mathrm{TiTv}}(a) = \begin{cases} 0 & \text{if } X_{ia} = X_{ja} \text{ and Ti/Tv} \\ 1/4 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Ti} \\ 1/2 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Tv} \\ 3/4 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Ti} \\ 1 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Tv} \end{cases} \tag{105}$$

With any of the three metrics given by Eqs. 103 - 105, we compute the pairwise distance between two instances i and j using Eq. 1 with $q = 1$. Assuming that all data entries X_{ia} are independent and identically distributed, we have already shown that the distribution of pairwise distances is asymptotically normal regardless of data distribution and value of q. Therefore, the distance distributions induced by each of the GWAS metrics given by Eqs. 103 - 105 are asymptotically normal. Thus, we will proceed by deriving the mean and variance for each distance distribution induced by these three GWAS metrics.

3.5.1 GM Distance Distribution

The expected value of the GM metric is given by the following.

$$\begin{aligned} \mathrm{E}\left[d_{ij}^{\mathrm{GM}}(a)\right] &= \sum_{k=0}^{1} k \cdot \mathrm{P}\left[d_{ij}^{\mathrm{GM}}(a) = k\right] \\ &= 0 \cdot \mathrm{P}\left[d_{ij}^{\mathrm{GM}}(a) = 0\right] + 1 \cdot \mathrm{P}\left[d_{ij}^{\mathrm{GM}}(a) = 1\right] \\ &= \mathrm{P}\left[d_{ij}^{\mathrm{GM}}(a) = 1\right] \\ &= 2\mathrm{P}[X_{ia} = 0, X_{ja} = 1] + 2\mathrm{P}[X_{ia} = 1, X_{ja} = 2] + 2\mathrm{P}[X_{ia} = 0, X_{ja} = 2] \\ &= 4(1-f_a)^3 f_a + 4(1-f_a) f_a^3 + 2(1-f_a)^2 f_a^2 \\ &= 2\left[2(1-f_a)^3 f_a + 2(1-f_a) f_a^3 + (1-f_a)^2 f_a^2\right] \\ &= 2F(a) \end{aligned} \tag{106}$$

where $F(a) = 2(1-f_a)^3 f_a + 2(1-f_a) f_a^3 + (1-f_a)^2 f_a^2$.

Then the expected pairwise GM distance between instances i and j is computed as follows.

$$\begin{aligned} \mathrm{E}\left(\sum_{a \in \mathcal{A}} d_{ij}^{\mathrm{GM}}(a)\right) &= \sum_{a \in \mathcal{A}} \mathrm{E}\left[d_{ij}^{\mathrm{GM}}(a)\right] \\ &= 2\sum_{a \in \mathcal{A}} F(a) \end{aligned} \tag{107}$$

The second moment about the origin for the GM distance is computed as follows.

$$\begin{aligned}
\mathrm{E}\left[(D_{ij})^2\right] &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}} d_{ij}^{\mathrm{GM}}(a)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a\in\mathcal{A}} \left(d_{ij}^{\mathrm{GM}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r\in\mathcal{A}}\sum_{s\leq r-1} d_{ij}^{\mathrm{GM}}(r)\cdot d_{ij}^{\mathrm{GM}}(s)\right] \\
&= \sum_{a\in\mathcal{A}}\left(\sum_{k=0}^{1} k^2\cdot \mathrm{P}\left[d_{ij}^{\mathrm{GM}}(a)=k\right]\right) \\
&+ 2\sum_{a\in\mathcal{A}}\sum_{s\leq r-1}\left(\sum_{k=0}^{1} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{GM}}(r)=k\right]\right)\cdot\left(\sum_{k=0}^{1} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{GM}}(s)=k\right]\right) \\
&= 2\sum_{a\in\mathcal{A}} F(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F(\lambda)
\end{aligned} \tag{108}$$

where $F(a) = 2(1-f_a)^3 f_a + 2(1-f_a)f_a^3 + (1-f_a)^2 f_a^2$.

Using the moments given by Eqs. 107 and 108, the variance is computed as follows.

$$\begin{aligned}
\mathrm{Var}(D_{ij}) &= \mathrm{E}\left[(D_{ij})^2\right] - \left[\mathrm{E}(D_{ij})\right]^2 \\
&= 2\sum_{a\in\mathcal{A}} F(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F(\lambda) - 4\left(\sum_{a\in\mathcal{A}}\right)^2 \\
&= 2\sum_{a\in\mathcal{A}} F(a) - 4\sum_{a\in\mathcal{A}} F^2(a) \\
&= 2\sum_{a\in\mathcal{A}} F(a)[1-2F(a)]
\end{aligned} \tag{109}$$

where $F(a) = 2(1-f_a)^3 f_a + 2(1-f_a)f_a^3 + (1-f_a)^2 f_a^2$.

With the mean and variance estimates given by Eqs. 107 and 109, the asymptotic GM distance distribution is given by the following.

$$D_{ij} \dot{\sim} \mathcal{N}\left(2\sum_{a\in\mathcal{A}} F(a), 2\sum_{a\in\mathcal{A}} F(a)[1-2F(a)]\right) \tag{110}$$

where $F(a) = 2(1-f_a)^3 f_a + 2(1-f_a)f_a^3 + (1-f_a)^2 f_a^2$.

3.5.2 AM Distance Distribution

The expected value of the AM metric is given by the following.

$$
\begin{aligned}
\mathrm{E}\left[d_{ij}^{\mathrm{AM}}(a)\right] &= \sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{AM}}(a)=k\right] \\
&= 0\cdot \mathrm{P}\left[d_{ij}^{\mathrm{AM}}(a)=0\right] + \frac{1}{2}\cdot \mathrm{P}\left[d_{ij}^{\mathrm{AM}}(a)=\frac{1}{2}\right] + 1\cdot \mathrm{P}\left[d_{ij}^{\mathrm{AM}}(a)=1\right] \\
&= \frac{1}{2}\left(2\mathrm{P}\left[X_{ia}=0, X_{ja}=1\right] + 2\mathrm{P}\left[X_{ia}=1, X_{ja}=2\right]\right) \\
&+ 2\mathrm{P}\left[X_{ia}=0, X_{ja}=2\right] \\
&= \mathrm{P}\left[X_{ia}=0, X_{ja}=1\right] + \mathrm{P}\left[X_{ia}=1, X_{ja}=2\right] + 2\mathrm{P}\left[X_{ia}=0, X_{ja}=2\right] \\
&= 2(1-f_a)^3 f_a + 2(1-f_a) f_a^3 + 2(1-f_a)^2 f_a^2 \\
&= 2\left[(1-f_a)^3 f_a + (1-f_a) f_a^3 + (1-f_a)^2 f_a^2\right] \\
&= 2F(a)
\end{aligned}
\tag{111}
$$

where $F(a) = (1-f_a)^3 f_a + (1-f_a) f_a^3 + (1-f_a)^2 f_a^2$ and $\mathcal{D} = \{0, \frac{1}{2}, 1\}$.

Then the expected pairwise AM distance between instances i and j is computed as follows.

$$
\begin{aligned}
\mathrm{E}\left(\sum_{a\in\mathcal{A}} d_{ij}^{\mathrm{AM}}(a)\right) &= \sum_{a\in\mathcal{A}} \mathrm{E}\left[d_{ij}^{\mathrm{AM}}(a)\right] \\
&= 2\sum_{a\in\mathcal{A}} F(a)
\end{aligned}
\tag{112}
$$

The second moment about the origin for the AM distance is computed as follows.

$$
\begin{aligned}
\mathrm{E}\left[(D_{ij})^2\right] &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}} d_{ij}^{\mathrm{AM}}(a)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a\in\mathcal{A}} \left(d_{ij}^{\mathrm{AM}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r\in\mathcal{A}}\sum_{s\le r-1} d_{ij}^{\mathrm{AM}}(r)\cdot d_{ij}^{\mathrm{AM}}(s)\right] \\
&= \sum_{a\in\mathcal{A}}\left(\sum_{k\in\mathcal{D}} k^2\cdot \mathrm{P}\left[d_{ij}^{\mathrm{AM}}(a)=k\right]\right) \\
&+ 2\sum_{a\in\mathcal{A}}\sum_{s\le r-1}\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{AM}}(r)=k\right]\right)\cdot\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{AM}}(s)=k\right]\right) \\
&= \sum_{a\in\mathcal{A}} G(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\le r-1}\prod_{\lambda\in\{r,s\}} F(\lambda)
\end{aligned}
\tag{113}
$$

where $G(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$ and $F(\lambda) = (1-f_\lambda)^3 f_\lambda + f_\lambda^3(1-f_\lambda) + (1-f_\lambda)^2 f_\lambda^2$.

Using the moments given by Eqs. 112 and 113, the variance is computed as follows.

$$
\begin{aligned}
\mathrm{Var}(D_{ij}) &= \mathrm{E}\left[(D_{ij})^2\right] - \left[\mathrm{E}(D_{ij})\right]^2 \\
&= \sum_{a\in\mathcal{A}} G(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F(\lambda) - 4\left(\sum_{a\in\mathcal{A}}\right)^2 \\
&= \sum_{a\in\mathcal{A}} G(a) - 4\sum_{a\in\mathcal{A}} F^2(a) \\
&= \sum_{a\in\mathcal{A}} \left[G(a) - 4F^2(a)\right]
\end{aligned}
\tag{114}
$$

where $G(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$ and $F(\lambda) = (1-f_\lambda)^3 f_\lambda + f_\lambda^3(1-f_\lambda) + (1-f_\lambda)^2 f_\lambda^2$.

With the mean and variance estimates given by Eqs. 112 and 114, the asymptotic AM distance distribution is given by the following.

$$
D_{ij} \dot{\sim} \mathcal{N}\left(2\sum_{a\in\mathcal{A}} F(a), \sum_{a\in\mathcal{A}}\left[G(a) - 4F^2(a)\right]\right) \tag{115}
$$

where $G(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$ and $F(\lambda) = (1-f_\lambda)^3 f_\lambda + f_\lambda^3(1-f_\lambda) + (1-f_\lambda)^2 f_\lambda^2$.

3.5.3 TiTv Distance Distribution

The TiTv metric allows for one to account for both genotype mismatch, allele mismatch, transition, and transversion. However, this added dimension of information requires knowledge of the nucleotide makeup at a particular locus. A sufficient conditions to compute the TiTv metric between instances i an j is that we know whether the nucleotides associated with a particular locus a are both purines (PuPu), purine and pyrimidine (PuPy), or both pyrimidines (PyPy). This information is always given in a particular data set. Let γ_0, γ_1, and γ_2 denote the probabilities of PuPu, PuPy, and PyPy, respectively, for the p loci of data matrix X. In real data, there are approximately twice as many transitions as there are transversions. Thas is, the probability of a transition P(Ti) is approximately twice the probability of transversion P(Tv). In order to enforce this in simulated data, we sample γ_0, γ_1, and γ_2 subject to the following constraints.

$$
\gamma_0 + \gamma_1 + \gamma_2 = 1 \tag{116}
$$

$$
\mathrm{P(Ti)} - 2\mathrm{P(Tv)} = 0 \tag{117}
$$

$$
1 - \mathrm{P(Ti)} - \mathrm{P(Tv)} = 0 \tag{118}
$$

Let y represent a random sample of size p from $\{\mathrm{PuPu}, \mathrm{PuPy}, \mathrm{PyPy}\}$. Let Y be a matrix with the following entries.

$$
Y_{ia} = \begin{cases} 0 & \text{if } X_{ia} \text{ is PuPu} \\ 1 & \text{if } X_{ia} \text{ is PuPy} \\ 2 & \text{if } X_{ia} \text{ is PyPy} \end{cases} \tag{119}
$$

Using the PuPu, PuPy, and PyPy encoding given previously, we compute the probability of a transversion occuring at a position X_{ia} as follows.

$$
\begin{aligned}
\mathrm{P}(\mathrm{Tv}) &= \mathrm{P}\left[Y_{ia}=1, Y_{ja}=0\right] + \mathrm{P}\left[Y_{ia}=0, Y_{ja}=1\right] \\
&\quad + \mathrm{P}\left[Y_{ia}=2, Y_{ja}=1\right] + \mathrm{P}\left[Y_{ia}=1, Y_{ja}=2\right] \\
&\quad + \mathrm{P}\left[Y_{ia}=2, Y_{ja}=0\right] + \mathrm{P}\left[Y_{ia}=0, Y_{ja}=2\right] \\
&= \gamma_0\gamma_1 + \gamma_1\gamma_0 + \gamma_1\gamma_2 + \gamma_2\gamma_1 + \gamma_0\gamma_2 + \gamma_2\gamma_0 \\
&= 2(\gamma_0\gamma_1 + \gamma_1\gamma_2 + \gamma_0\gamma_2)
\end{aligned} \tag{120}
$$

Using the constraint given by Eq. 117, the probability of a transition occuring at a position X_{ia} is computed as follows.

$$
\begin{aligned}
\mathrm{P}(\mathrm{Ti}) &= 1 - \mathrm{P}(\mathrm{Tv}) \\
&= 1 - 2(\gamma_0\gamma_1 + \gamma_1\gamma_2 + \gamma_0\gamma_2)
\end{aligned} \tag{121}
$$

Based on the constraint given by Eq. 118, it is clear that we have $\mathrm{P}(\mathrm{Tv}) = \frac{1}{3}$. This implies that $\mathrm{P}(\mathrm{Ti}) = \frac{2}{3}$. In order to satisfy the constraint given by Eq. 116, we can sample γ_0 so that we have the following.

$$
\begin{aligned}
\gamma_2 &= \frac{6(1-\gamma_0) + 2\sqrt{9(1-\gamma_0)^2 - 6(6\gamma_0^2 - 6\gamma_0 + 1)}}{12} \\
\gamma_1 &= 1 - \gamma_0 - \gamma_2 \\
0 &< \gamma_1 < 1 \\
0 &< \gamma_2 < 1
\end{aligned} \tag{122}
$$

The first equality given in Eq. 122 comes from the fact that we have a known probability γ_0, which gives two equations in two unknowns (γ_1 and γ_2) that come from the constraint given by Eq. 116 and the fact that $\mathrm{P}(\mathrm{Tv}) = 2(\gamma_0\gamma_1 + \gamma_1\gamma_2 + \gamma_0\gamma_2) = \frac{1}{3}$. The solution $(\gamma_0, \gamma_1, \gamma_2)$ to the constraint equations 122 allows one to simulate a SNP data set that approximately satisfies constraint 117.

We proceed by computing $\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a) = k\right]$ for each $k \in \mathcal{D} = \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}$. We derive $\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a) = 0\right]$ as follows.

$$
\begin{aligned}
\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a) = 0\right] &= \mathrm{P}\left[y_a = 0, X_{ia} = X_{ja}\right] \\
&\quad + \mathrm{P}\left[y_a = 1, X_{ia} = X_{ja}\right] \\
&\quad + \mathrm{P}\left[y_a = 2, X_{ia} = X_{ja}\right] \\
&= \gamma_0\left[(1-f_a)^2 + 2f_a(1-f_a) + f_a^2\right] \\
&\quad + \gamma_1\left[(1-f_a)^2 + 2f_a(1-f_a) + f_a^2\right] \\
&\quad + \gamma_2\left[(1-f_a)^2 + 2f_a(1-f_a) + f_a^2\right] \\
&= (\gamma_0 + \gamma_1 + \gamma_2)\left[(1-f_a)^2 + 2f_a(1-f_a) + f_a^2\right] \\
&= (1-f_a)^2 + 2f_a(1-f_a) + f_a^2
\end{aligned} \tag{123}
$$

We derive $\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a) = \frac{1}{4}\right]$ as follows.

$$
\begin{aligned}
\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=\frac{1}{4}\right] &= 2\mathrm{P}\left[y_a=0, X_{ia}=0, X_{ja}=1\right] \\
&+ 2\mathrm{P}\left[y_a=0, X_{ia}=1, X_{ja}=2\right] \\
&+ 2\mathrm{P}\left[y_a=2, X_{ia}=0, X_{ja}=1\right] \\
&+ 2\mathrm{P}\left[y_a=2, X_{ia}=1, X_{ja}=2\right] \\
&= 4\gamma_0(1-f_a)^3 f_a + 4\gamma_0 f_a^3(1-f_a) + 4\gamma_2(1-f_a)^3 f_a \\
&+ 4\gamma_2 f_a^3(1-f_a) \\
&= 4\gamma_0\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] \\
&+ 4\gamma_2\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] \\
&= 4(\gamma_0+\gamma_2)\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right]
\end{aligned}
\tag{124}
$$

We derive $\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=\frac{1}{2}\right]$ as follows.

$$
\begin{aligned}
\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=\frac{1}{2}\right] &= 2\mathrm{P}\left[y_a=1, X_{ia}=0, X_{ja}=1\right] \\
&+ 2\mathrm{P}\left[y_a=1, X_{ia}=1, X_{ja}=2\right] \\
&= 4\gamma_1(1-f_a)^3 f_a + 4\gamma_1 f_a^3(1-f_a) \\
&= 4\gamma_1\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right]
\end{aligned}
\tag{125}
$$

We derive $\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=\frac{3}{4}\right]$ as follows.

$$
\begin{aligned}
\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=\frac{3}{4}\right] &= 2\mathrm{P}\left[y_a=0, X_{ia}=0, X_{ja}=2\right] \\
&+ 2\mathrm{P}\left[y_a=2, X_{ia}=0, X_{ja}=2\right] \\
&= 2\gamma_0(1-f_a)^2 f_a^2 + 2\gamma_2(1-f_a)^2 f_a^2 \\
&= 2(\gamma_0+\gamma_2)(1-f_a)^2 f_a^2
\end{aligned}
\tag{126}
$$

We derive $\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=1\right]$ as follows.

$$
\begin{aligned}
\mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=1\right] &= 2\mathrm{P}\left[y_a=1, X_{ia}=0, X_{ja}=2\right] \\
&= 2\gamma_1(1-f_a)^2 f_a^2
\end{aligned}
\tag{127}
$$

Using Eqs. 123 - 127, we compute the expected TiTv distance between instances i and j as follows.

$$
\begin{aligned}
\mathrm{E}(D_{ij}) &= \sum_{a\in\mathcal{A}}\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=k\right]\right) \\
&= (\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}}\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] \\
&+ \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}}(1-f_a)^2 f_a^2 \\
&= (\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a)
\end{aligned}
\tag{128}
$$

where $F(a)=(1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G(a)=(1-f_a)^2 f_a^2$.

The second moment about the origin for the TiTv distance is computed as follows.

$$
\begin{aligned}
\mathrm{E}\left[(D_{ij})^2\right] &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}} d_{ij}^{\mathrm{TiTv}}(a)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a\in\mathcal{A}}\left(d_{ij}^{\mathrm{TiTv}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r\in\mathcal{A}}\sum_{s\le r-1} d_{ij}^{\mathrm{TiTv}}(r)\cdot d_{ij}^{\mathrm{TiTv}}(s)\right] \\
&= \sum_{a\in\mathcal{A}}\left(\sum_{k\in\mathcal{D}} k^2\cdot \mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(a)=k\right]\right) \\
&+ 2\sum_{a\in\mathcal{A}}\sum_{s\le r-1}\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(r)=k\right]\right)\cdot\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[d_{ij}^{\mathrm{TiTv}}(s)=k\right]\right) \\
&= \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a) \\
&+ 2\sum_{r\in\mathcal{A}}\sum_{s\le r-1}\prod_{\lambda\in\{r,s\}}\left([\gamma_0+\gamma_2+2\gamma_1]F(\lambda)+\left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G(\lambda)\right)
\end{aligned}
\tag{129}
$$

where $F(\lambda) = (1-f_\lambda)^3 f_\lambda + f_\lambda^3(1-f_\lambda)$ and $G(\lambda) = (1-f_\lambda)^2 f_\lambda^2$.

Using the moments given by Eqs. 128 and 129, the variance is computed as follows.

$$
\begin{aligned}
\mathrm{Var}(D_{ij}) &= \mathrm{E}\left[(D_{ij})^2\right] - \left[\mathrm{E}(D_{ij})\right]^2 \\
&= \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a) \\
&+ 2\sum_{r\in\mathcal{A}}\sum_{s\le r-1}\prod_{\lambda\in\{r,s\}}\left([\gamma_0+\gamma_2+2\gamma_1]F(\lambda)+\left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G(\lambda)\right) \\
&- \left([\gamma_0+\gamma_2+2\gamma_1]\sum_{a\in\mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a)\right)^2 \\
&= \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a) \\
&+ \sum_{a\in\mathcal{A}}\left([\gamma_0+\gamma_2+2\gamma_1]F(a)+\left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G(a)\right)^2
\end{aligned}
\tag{130}
$$

where $F(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G(a) = (1-f_a)^2 f_a^2$.

With the mean and variance estimates given by Eqs. 128 and 130, the asymptotic TiTv distance distribution is given by the following.

$$
\begin{aligned}
D_{ij} \dot{\sim} \mathcal{N}\Bigg(& (\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a), \\
& \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a) \\
& + \sum_{a\in\mathcal{A}}\left([\gamma_0+\gamma_2+2\gamma_1]F(a)+\left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G(a)\right)^2\Bigg)
\end{aligned}
\tag{131}
$$

where $F(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G(a) = (1-f_a)^2 f_a^2$.

3.6 Resting-State fMRI Distance Distribution

For resting-state fMRI (rs-fMRI), the data consists of correlation matrices for each instance. These correlations are between different ROIs for a particular brain atlas. We would like the attributes to be the ROIs themselves, which leads us to the following metric.

$$d_{ij}^{\mathrm{ROI}}(a) = \sum_{k \neq a} \left| A_{ka}^{(i)} - A_{ka}^{(j)} \right| \tag{132}$$

where $A_{ka}^{(i)}$ and $A_{ka}^{(j)}$ are the correlations between ROI a and ROI k for instances i and j, respectively. In order for comparisons between different correlations to be possible, we first perform a Fisher r-to-z transform on the correlations. We then load all of the transformed correlations into a $p(p-1) \times m$ matrix X with the following form.

$$X = \begin{bmatrix} \hat{A}_{12}^{(1)} & \hat{A}_{12}^{(2)} & \hat{A}_{12}^{(3)} & \dots & \hat{A}_{12}^{(m)} \\ \hat{A}_{13}^{(1)} & \hat{A}_{13}^{(2)} & \hat{A}_{13}^{(3)} & \dots & \hat{A}_{13}^{(m)} \\ \hat{A}_{14}^{(1)} & \hat{A}_{14}^{(2)} & \hat{A}_{14}^{(3)} & \dots & \hat{A}_{14}^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{A}_{1p}^{(1)} & \hat{A}_{1p}^{(2)} & \hat{A}_{1p}^{(3)} & \dots & \hat{A}_{1p}^{(m)} \\ \hat{A}_{21}^{(1)} & \hat{A}_{21}^{(2)} & \hat{A}_{21}^{(3)} & \dots & \hat{A}_{21}^{(m)} \\ \hat{A}_{23}^{(1)} & \hat{A}_{23}^{(2)} & \hat{A}_{23}^{(3)} & \dots & \hat{A}_{23}^{(m)} \\ \hat{A}_{24}^{(1)} & \hat{A}_{24}^{(2)} & \hat{A}_{24}^{(3)} & \dots & \hat{A}_{24}^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{A}_{2p}^{(1)} & \hat{A}_{2p}^{(2)} & \hat{A}_{2p}^{(3)} & \dots & \hat{A}_{2p}^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{A}_{p1}^{(1)} & \hat{A}_{p1}^{(2)} & \hat{A}_{p1}^{(3)} & \dots & \hat{A}_{p1}^{(m)} \\ \hat{A}_{p2}^{(1)} & \hat{A}_{p2}^{(2)} & \hat{A}_{p2}^{(3)} & \dots & \hat{A}_{p2}^{(m)} \\ \hat{A}_{p3}^{(1)} & \hat{A}_{p3}^{(2)} & \hat{A}_{p3}^{(3)} & \dots & \hat{A}_{p3}^{(m)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{A}_{p,(p-1)}^{(1)} & \hat{A}_{p,(p-1)}^{(2)} & \hat{A}_{p,(p-1)}^{(3)} & \dots & \hat{A}_{p,(p-1)}^{(m)} \end{bmatrix} \tag{133}$$

where $\hat{A}_{ka}^{(i)}$ is the r-to-z transformed correlation between ROIs a and k for instance i.

We further transform the data matrix X by standardizing so that each of the m columns has zero mean and unit variance. Therefore, the data in matrix X are standard normal. Recall from Eqs. 49 and 50, that the mean and variance of the Manhattan ($q = 1$) distance distribution for standard normal data are $\frac{2p}{\sqrt{\pi}}$ and $\frac{2(\pi-2)p}{\pi}$, respectively. This allows us to easily derive the expected pairwise distance between instances i and j in rs-fMRI data as follows.

$$
\begin{aligned}
\mathrm{E}(D_{ij}) &= \mathrm{E}\left(\sum_{a\in\mathcal{A}}\sum_{k\neq a}\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|\right)\\
&= \sum_{a\in\mathcal{A}}\sum_{k\neq a}\mathrm{E}\left(\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|\right)\\
&= \sum_{a\in\mathcal{A}}\sum_{k\neq a}\frac{2}{\sqrt{\pi}}\\
&= \frac{2p(p-1)}{\sqrt{\pi}}
\end{aligned}
\tag{134}
$$

Due to the dependencies that exist between terms in the double sum when computing the rs-fMRI distance, linearity no longer applies to the variance operator. We proceed by writing the form of the variance as follows.

$$
\begin{aligned}
\mathrm{Var}(D_{ij}) &= \mathrm{Var}\left(\sum_{a\in\mathcal{A}}\sum_{k\neq a}\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|\right)\\
&= \sum_{a=1}^{p-1}\mathrm{Var}\left(\sum_{k=a+1}^{p}2\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|\right)\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|,\ \sum_{s=r+1}^{p}2\left|\hat{A}^{(i)}_{rs}-\hat{A}^{(j)}_{rs}\right|\right)\\
&= \sum_{a=1}^{p-1}\sum_{k=a+1}^{p}\mathrm{Var}\left(2\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|\right)\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|,\ \sum_{s=r+1}^{p}2\left|\hat{A}^{(i)}_{rs}-\hat{A}^{(j)}_{rs}\right|\right)\\
&= \sum_{a=1}^{p-1}\sum_{k=a+1}^{p-1}\frac{4(\pi-2)}{\pi}\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|,\ \sum_{s=r+1}^{p}2\left|\hat{A}^{(i)}_{rs}-\hat{A}^{(j)}_{rs}\right|\right)\\
&= \frac{2p(\pi-2)(p-1)}{\pi}\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}^{(i)}_{ak}-\hat{A}^{(j)}_{ak}\right|,\ \sum_{s=r+1}^{p}2\left|\hat{A}^{(i)}_{rs}-\hat{A}^{(j)}_{rs}\right|\right)
\end{aligned}
\tag{135}
$$

In order to have a formula in terms of the number of ROIs p only, we must estimate the double sum on the right-hand side of Eq. 135. Through simulation, it can be seen that the difference between the sample variance $S^2_{D_{ij}}$ and $\frac{2p(\pi-2)(p-1)}{\pi}$ has a quadratic relationship with p. More explicitly, we have the following relationship.

$$
S^2_{D_{ij}} - \frac{2p(\pi-2)(p-1)}{\pi} = \beta_1 p^2 + \beta_0 p \tag{136}
$$

The coefficient estimates found through least squares fitting are $\beta_0 = -\beta_1 \approx 0.08$. These estimates allow one to infer a functional form for the double sum in the right-hand

side of Eq. 135 that is actually proportional to $\frac{2p(\pi-2)(p-1)}{\pi}$. That is, we have the following formula for approximating the double sum.

$$2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\text{Cov}\left(\sum_{k=a+1}^{p}2|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}|,\sum_{s=r+1}^{p}2|\hat{A}_{rs}^{(i)}-\hat{A}_{rs}^{(j)}|\right)=\frac{p(\pi-2)(p-1)}{4\pi} \quad (137)$$

Therefore, the variance of the rs-fMRI distances is approximated well by the following.

$$\text{Var}(D_{ij})=\frac{9p(\pi-2)(p-1)}{4\pi} \quad (138)$$

With the mean and variance estimates given by Eqs. 134 and 138, we have the following asymptotic distribution for rs-fMRI distances.

$$D_{ij}\dot{\sim}\mathcal{N}\left(\frac{2p(p-1)}{\sqrt{\pi}},\frac{9p(\pi-2)(p-1)}{4\pi}\right) \quad (139)$$

Consider the max-min normalized rs-fMRI distance given by the following equation.

$$D_{ij}^{*}=\sum_{a\in\mathcal{A}}\sum_{k\neq a}\frac{|A_{ak}^{(i)}-A_{ak}^{(j)}|}{\max(a)-\min(a)} \quad (140)$$

Assuming that the data X has been r-to-z transformed and standardized, we can easily compute the expected attribute range and variance of the attribute range. The expected maximum of a given attribute in data matrix X is estimated by the following.

$$\text{E}\left(X_a^{\alpha}-X_a^{\omega}\right)=2\mu_{\alpha}^{(1)}(m,p)=2\left[\frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m(p-1)}\right)}-\Phi^{-1}\left(\frac{1}{m(p-1)}\right)\right] \quad (141)$$

The variance can be esimated with the following.

$$\text{Var}\left(X_a^{\alpha}-X_a^{\omega}\right)=\frac{\pi^2}{6\log[m(p-1)]} \quad (142)$$

Let $\mu_{D_{ij}}$ and $\sigma_{D_{ij}}^2$ denote the mean and variance of the rs-fMRI distance distribution given by Eqs. 134 and 138. Using the formulas for the mean and variance of the max-min normalized distance distribution given in Eq. 93, we have the following asymptotic distribution for the max-min normalized rs-fMRI distances.

$$D_{ij}^{*}\dot{\sim}\mathcal{N}\left(\frac{\mu_{D_{ij}}}{2\mu_{\alpha}^{(1)}(m,p)},\frac{6\sigma_{D_{ij}}^2\log[m(p-1)]}{\pi^2+24\left[\mu_{\alpha}^{(1)}(m,p)\right]^2\log[m(p-1)]}\right) \quad (143)$$

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.

Table 3. Summary of distance distribution derivations for standard normal and standard uniform data.

q-Metric	Data	Stat	Formula (Eq. #)
standard (**Eq. 2**)	$\mathcal{N}(0,1)$	mean	$\left(\frac{2^q\Gamma\left(\frac{q+1}{2}\right)p}{\sqrt{\pi}}\right)^{1/q}$ (**38**)
	$\mathcal{N}(0,1)$	variance	$\frac{4^q p}{q^2\left(\frac{2^q\Gamma\left(\frac{1}{2}q+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}}-\frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]$ (**38**)
	$\mathcal{U}(0,1)$	mean	$\left(\frac{2p}{(q+2)(q+1)}\right)^{1/q}$ (**48**)
	$\mathcal{U}(0,1)$	variance	$\frac{p}{q^2\left(\frac{2p}{(q+2)(q+1)}\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{1}{(q+1)(2q+1)}-\left(\frac{2}{(q+2)(q+1)}\right)^2\right]$ (**48**)
max-min normalized (**Eq. 59**)	$\mathcal{N}(0,1)$	mean	$\boxed{\frac{\mu_{D_{ij}^{(q)}}}{2\mu_\alpha^{(1)}(m)}}$ (**93**) where $\mu_{D_{ij}^{(q)}}$ and $\mu_\alpha^{(1)}(m)$ are given by Eqs. 38 and 87, respectively.
	$\mathcal{N}(0,1)$	variance	$\boxed{\frac{6\log(m)\sigma^2_{D_{ij}^{(q)}}}{\pi^2+24\left[\mu_\alpha^{(1)}(m)\right]^2\log(m)}}$ (**93**) where $\sigma^2_{D_{ij}^{(q)}}$ and $\mu_\alpha^{(1)}(m)$ are given by Eqs. 38 and 87, respectively.
	$\mathcal{U}(0,1)$	mean	$\boxed{\frac{(m+1)\mu_{D_{ij}^{(q)}}}{m-1}}$ (**101**) where $\mu_{D_{ij}^{(q)}}$ is given by Eq. 48
	$\mathcal{U}(0,1)$	variance	$\boxed{\frac{(m+2)(m+1)^2\sigma^2_{D_{ij}^{(q)}}}{m^3-m+2}}$ (**101**) where $\sigma^2_{D_{ij}^{(q)}}$ is given by Eq. 48

Table 4. Summary of distance distribution derivations for GWAS data.

GWAS-Metric	Stat	Formula (Eq. #)
GM (**Eq. 103**)	mean	$\boxed{2\sum_{a\in\mathcal{A}} F(a)}$ (**110**) where $F(a) = 2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2$
	variance	$\boxed{2\sum_{a\in\mathcal{A}} F(a)[1-2F(a)]}$ (**110**) where $F(a) = 2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2$
AM (**Eq. 104**)	mean	$\boxed{2\sum_{a\in\mathcal{A}} F(a)}$ (**115**) where $F(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + (1-f_a)^2 f_a^2$
	variance	$\boxed{\sum_{a\in\mathcal{A}}[G(a) - 4F^2(a)]}$ (**115**) where $F(a) = 2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2$ and $G(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$
TiTv (**Eq. 105**)	mean	$\boxed{(\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a)}$ (**131**) where $F(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G(a) = (1-f_a)^2 f_a^2$
	mean	$\boxed{\left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G(a) + \sum_{a\in\mathcal{A}}\left([\gamma_0+\gamma_2+2\gamma_1]F(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G(a)\right)^2}$ (**131**) where $F(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G(a) = (1-f_a)^2 f_a^2$

Table 5. Summary of distance distribution derivations for rs-fMRI data.

rs-fMRI - Metric	Stat	Formula (Eq. #)
standard (**Eq. 132**)	mean	$\frac{2p(p-1)}{\sqrt{\pi}}$ (**139**)
	variance	$\frac{9p(\pi-2)(p-1)}{4\pi}$ (**139**)
max-min normalized (**Eq. 140**)	mean	$\boxed{\frac{\mu_{D_{ij}}}{2\mu_\alpha^{(1)}(m,p)}}$ (**143**) where $\mu_{D_{ij}}$ and $\mu_\alpha^{(1)}(m,p)$ are given by Eqs. 140 and 142
	variance	$\boxed{\frac{6\sigma^2_{D_{ij}}\log[m(p-1)]}{\pi^2+24\left[\mu_\alpha^{(1)}(m,p)\right]^2\log[m(p-1)]}}$ (**143**) where $\sigma^2_{D_{ij}}$ and $\mu_\alpha^{(1)}(m,p)$ are given by Eqs. 140 and 142

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).

6 Data simulations

Each simulated data set $X^{(m\times p)}$ contains m instances and p features, where

$$m \in \{100, 250, 500\} \text{ and} \tag{144}$$

$$p \in \{1000, 2000, 3000, 4000, 5000\}. \tag{145}$$

All combinations of m and p were explored in order to determine how neighborhood selection parameters change with dimensionality. We considered a balanced binary outcome only so that there were exactly $m/2$ cases and $m/2$ controls. In each simulation, 10% of the total number of features p were functionally related to the outcome variable while the remaining 90% were simply background features with no effect. Functional features were given either main or interaction effect in order to create a random mixed effects data set for which optimal neighborhood method parameters could be calculated.

6.1 Interaction effects

We extend the interaction effect simulation introduced in [11]. This method first generates a random graph from either Erdős-Rényi or Scale-free degree distribution. In the control group, functional features are given large pairwise correlations with all other features. Differential correlations (interaction effects) are created by randomly permuting functional feature data entries within the case group only, which destroys and preserves the correlation in case and control groups, respectively. This method creates large effect sizes, which are easily detected by nearest-neighbor distance based methods. The reason for this ease of detection is the uniformity in low and high correlations in case and control groups, respectively. In order to establish more influence over the number of differential pairwise correlations, we simulated correlation matrices for case and control groups directly. We allowed only functional connections to be given differential correlation between case and control groups, where a functional connection is simply the presence of an edge (or link) from one feature to a functional feature in the random network that is generated.

Analogous to the interaction simulation method given in [11], we give the control group high and low positive pairwise correlation between network connected and non-connected features, respectively. The average magnitudes of the high and low correlations in the control group are determined by fixed parameters ρ^{hi} and ρ^{lo}, respectively. For all simulations, we fixed the low correlation parameter ρ^{lo}. In order to examine how interaction effect size changes with functional connection pairwise correlation, we let $\rho^{\text{hi}} \in \{0.2, 0.5, 0.8\}$. Interaction effect size increases and decreases monotonically as ρ^{hi} increases and decreases, respectively. Interaction effect size is highly related to how connected a functional feature is in the random network. We control network connectivity by the probability of edge inclusion for Erdős-Rényi networks and fixed node degree for Scale-free networks. Similar to ρ^{hi}, interaction effect size increases and decreases monotonically as functional network connectivity increases and decreases, respectively. Furthermore, we determine interaction effect sizes by giving functional connection pairwise correlations in the case group values determined by the parameter b^{int}, which is defined as

$$b^{\text{int}} = -t\rho^{\text{hi}} + (1-t)\rho^{\text{hi}}, \text{ where } t \in [0,1]. \tag{146}$$

As $t \to 0$, the effect size decreases monotonically. On the other hand, the effect size increases monotonically as $t \to 1$. By controlling ρ^{hi}, b^{int}, and the level of network connectivity, we have the ability to more finely control the interaction effect size than the method presented in [11].

After correlation matrices are created for cases and controls, we compute the upper triangular Cholesky factors (U^{case} and U^{ctrl}) for each correlation matrix. We simulate null data matrices X^{case} and X^{ctrl} for cases and controls, respectively, such that

$$x_{ij}^{\text{case}}, x_{ij}^{\text{ctrl}} \sim \mathcal{N}(0,1) \quad \forall i, j. \tag{147}$$

Multiplication of X^{case} and X^{ctrl} by the Cholesky factors U^{case} and U^{ctrl}, respectively, produce case and control sub-matrices Y^{case} and Y^{ctrl} with the correlation structure described previously. These sub-matrices are then combined into a single data $m \times p$ matrix given by

$$X = \begin{bmatrix} Y^{\text{ctrl}} \\ \text{- - - -} \\ Y^{\text{case}} \end{bmatrix}. \tag{148}$$

A diagram outlining the interaction simulation algorithm for a data set consisting of 7 features is shown in Fig. 8. Boxes 1 and 2 display the random network and its characteristics, such as, adjacencies and node degrees. In particular, boxes 1 and 2 show

the features that are selected to be functional (highlighted in green). Box 3 shows the case and control correlation matrices generated using input parameters ρ^{hi}, ρ^{lo}, and b^{int}. In the control group, high correlation (red) is assigned to all connected pairs from the network. That is,

$$P_{ij}^{\text{ctrl}} = \rho^{\text{hi}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1). \tag{149}$$

In both case and control groups, low correlation (red) is assigned to non-connected features from the network. These low correlations are given by

$$P^{\text{case}} = P^{\text{ctrl}} = \rho^{\text{lo}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1). \tag{150}$$

In the case group, pairwise correlations associated with functional connections (blue) from the network are assigned a correlation that is functionally related (see Eq. 146) to high correlations in the control group. These correlations are given by

$$P_{ij}^{\text{case}} = b^{\text{int}} + \varepsilon_{ij}, \quad \varepsilon_{ij} \in \mathcal{N}(0,1). \tag{151}$$

Other than entries associated with functional connections, the case and control correlation matrices are identical. Box 4 shows the Cholesky decompositions for P^{ctrl} and P^{case}, which are given by

$$\begin{aligned} P^{\text{ctrl}} &= U^{\text{ctrl}} \left(U^{\text{ctrl}}\right)^{\text{T}} \quad \text{and} \\ P^{\text{case}} &= U^{\text{case}} \left(U^{\text{case}}\right)^{\text{T}}. \end{aligned} \tag{152}$$

Random case and control data with correlation structure determined by P^{case} and P^{ctrl}, respectively, are created as mentioned previously (box 5). These sub-matrices are given by

$$\begin{aligned} Y^{\text{ctrl}} &= X^{\text{ctrl}} \left(U^{\text{ctrl}}\right)^{\text{T}}, \quad x_{ij}^{\text{ctrl}} \sim \mathcal{N}(0,1) \quad \text{and} \\ Y^{\text{case}} &= X^{\text{case}} \left(U^{\text{case}}\right)^{\text{T}}, \quad x_{ij}^{\text{case}} \sim \mathcal{N}(0,1). \end{aligned} \tag{153}$$

The full data set, given previously by Eq. 148, concludes the generation of the full $m \times p$ data set X with interaction effects.

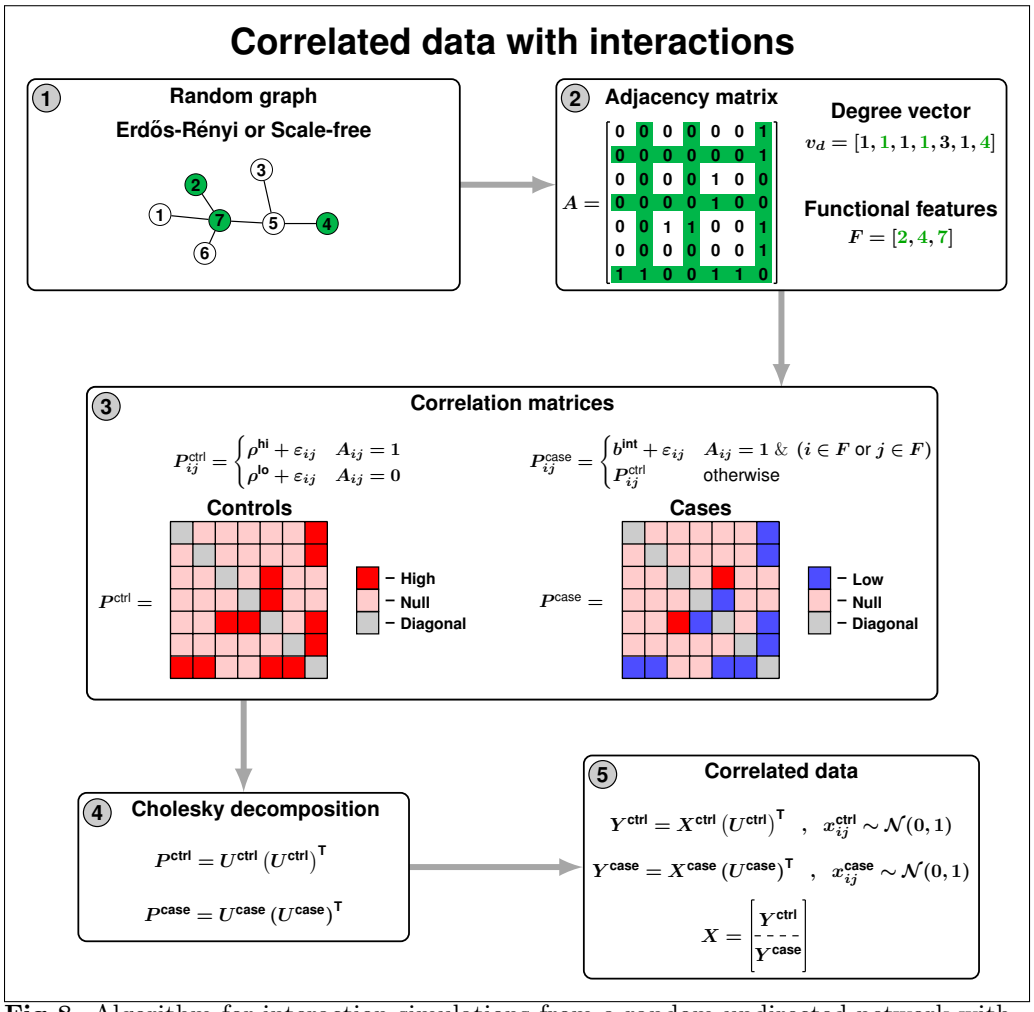

Fig 8. Algorithm for interaction simulations from a random undirected network with seven nodes (or features). **Box 1:** A random network is generated, whose degree distribution is either Erdős-Rényi or Scale-free. **Box 2:** Adjacency matrix (A) and degree vector (v_d) corresponding to the random network are computed and functional features (F) are randomly selected from those with positive degree. **Box 3:** Two correlation matrices are generated for cases and controls. In the control group, high (ρ^{hi}) and low (ρ^{lo}) correlations are assigned to connected ($A_{ij} = 1$) and non-connected ($A_{ij} = 0$) feature pairs, respectively. In the case group, differential correlation (b^{int}) is applied to functional connections. **Box 4:** Upper triangular Cholesky factors are computed for case/control correlation matrices. **Box 5:** Standard normal random data matrices (X^{ctrl} and X^{case}) are given correlation structure associated with case and control groups and combined into full data matrix with interaction effects (X).

6.2 Main effects

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. Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, NY, 2004.

4. Irwin Miller and Marylees Miller. *John E. Freund's Mathematical Statistics with Applications*. Pearson Prentice Hall, 7 edition, 2004.

5. Marko Robnik Šikonja and Igor Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 53:23 – 69, February 2003.

6. Ryan J. Urbanowicz, Melissa Meeker, William LaCava, Randal S. Olson, and Jason H. Moore. Relief-Based Feature Selection: Introduction and Review. *arXiv:1711.08421 [cs.DS]*, 2018.

7. E. J. Gumbel. The Distribution of the Range. *The Annals of Mathematical Statistics*, 18(3):384–412, September 1947.

8. Sourav Chatterjee. *Superconcentration and Related Topics*. 1439-7382. Springer International Publishing, 1 edition, 2014.

9. Harald Cramér. *Mathematical Methods of Statistics*, volume 1. Princeton University Press, reprint, revised edition, 1999.

10. M. Arabnejad, B. A. Dawkins, W. S. Bush, B. C. White, A. R. Harkness, and B. A. McKinney. Transition-transversion encoding and genetic relationship metfic in ReliefF feature selection improves pathway enrichment in GWAS. *BioData Mining*, 11(23), 2018.

11. 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.