

Novel metrics and nearest-neighbor distance distributions in high dimensional bioinformatics data

Bryan A. Dawkins¹, Trang T. Le² and Brett A. McKinney^{1,3,*}

¹Department of Mathematics, University of Tulsa, Tulsa, OK 74104, USA

²Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104

³Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104, USA.

Abstract

Nearest-neighbor projected distance regression (NPDR) is a feature selection algorithm that is able to detect interactions in high dimensional data. The performance of NPDR and other nearest neighbor methods depends on the metric for computing neighborhoods and the expected moments of the distribution of pairwise distances for the given data type. We derive general analytical expressions for distributional properties of pairwise distances for L_q metrics for Gaussian and uniform data with p attributes and m instances. These expressions are applicable to the analysis of gene expression data. We derive similar analytical expressions for a new metric for genome-wide association study data (categorical predictors) and a new metric for resting-state fMRI data (correlation-based predictors). In addition, we consider the effect of correlation in the data.

Author summary

Introduction

Feature selection that relies on nearest neighbor algorithms in order to determine relative feature importance requires an understanding of distributional properties for a variety of different metrics. This is, in large part, due to how various statistical effects change distance distributions. For continuous data, L_q metrics with $q = 1$ or $q = 2$ are those most commonly used in this context. For data from standard normal ($\mathcal{N}(0,1)$) or standard uniform ($\mathcal{U}(0,1)$) distributions, the asymptotic behavior of the L_q metrics is known. However, detailed derivations of these distance distribution asymptotics are not commonly found or mentioned in the literature on nearest-neighbor distance based feature selection [1–3]. Furthermore, there is much work to be done to better understand new metrics in discrete data, such as, genome-wide association studies (GWAS) data or correlation data like resting-state fMRI (rs-fMRI).

Much work has been done in feature selection for rs-fMRI data [4–7]. Typical feature selection methods include, but are not limited to, best subset feature selection, k-fold cross-validation, and nested cross-validation. In each method, a modeling procedure is chosen along with selected features to optimize some objective, such as, classification accuracy or mean squared error. The features to be selected are usually Regions of Interest (ROIs), which are formed by averaging the time series from highly correlated voxels. By combining voxels into a single ROI, the feature space is greatly reduced. Typically, correlations are then computed between all pairs of ROIs. A matrix of pairwise ROI-ROI correlations is created for each instance (or subject) in a data set. To the best of our knowledge, nearest-neighbor distance based feature selection has not been applied in the context of rs-fMRI. Since these nearest-neighbor distance-based methods have

been shown to be able to detect interactions in high-dimensional data [1, 2, 8], rs-fMRI data is potentially one area in which these methods have not sufficiently exploited. Therefore, we introduce a new metric to be used in combination with NPDR in order to explore potential insights these methods may provide in time series-correlation (ts-corr) based data like rs-fMRI. In this manuscript, we derive asymptotic estimates for the mean and variance of distance distributions induced by our new ts-corr based metric.

Newly introduced to feature selection in GWAS data is a metric that accounts for genotype mismatch (GM), allele mismatch (AM), transitions (Ti), and transversions (Tv) [9]. This TiTv metric provides one additional dimension of information for which GM and AM metrics do not account. Another positive aspect of this metric is its comparable simplicity to the GM and AM metrics. That is, it takes on a finite number of discrete values. We will derive asymptotic formulas for the mean and variance for all three of these GWAS metrics. Since the TiTv metric has been introduced only recently, all of our associated derivations will be new contributions.

Optimal choices of neighborhood selection parameters, such as, fixed-radius or fixed-k depend on distance distributional properties with respect to the instance dimension. As neighborhood order increases, nearest neighbor distance based algorithms get better at detecting main effects [8]. On the other hand, their ability to detect interaction effects decreases as neighborhood order increases [8]. These different statistical effects impact distance distributions by introducing positive skewness and increased variance, which can lead to changes in neighborhood inclusion. In order to understand how statistical effects impact distance distributions in continuous and discrete data types, we first derive distance asymptotics for null data where instances are independently and identically distributed and there is no correlation between features. Using these derivations, we can then determine how statistical effects and correlation change distance distributional properties from the null case.

We begin with derivations applicable to continuously distributed data sets with m instances and p features. From these more general derivations, we focus on the cases of standard normal and standard uniform data distributions. We then make a transition to discrete data in which each value in the $m \times p$ data matrix is from a binomial distribution parameterized by $n = 2$ trials and some success probability. The final set of asymptotic results will be for our ts-corr metric, with a particular emphasis on rs-fMRI data. Lastly, we show how correlation in the attribute space changes distance distributional properties.

1 Derivations of distance asymptotics for common metrics used in continuous 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

$$D_{ij}^{(q)} = \left(\sum_{a \in \mathcal{A}} |d_{ij}(a)|^q \right)^{1/q}, \quad (1)$$

which is typically Manhattan ($q = 1$) but may also be Euclidean ($q = 2$). The quantity $d_{ij}(a)$, 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 $d_{ij}(a)$ 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 (d^{num}) attribute a may be

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

where \hat{X} represents the standardized data matrix X . We use a simplified $d_{ij}(a)$ notation in place of the $\text{diff}(a, (i, j))$ 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 $d_{ij}^{\text{num}}(a)$ 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:

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

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

1.1 Distribution of pairwise distances

Suppose that $X_{ia}, X_{ja} \stackrel{iid}{\sim} \mathcal{F}_X(\mu_X, \sigma_X^2)$ 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

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

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

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

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 [10] can be applied to show that

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

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

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

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 of distance distributions for standard uniform data for Euclidean ($q = 2$) and Manhattan ($q = 1$) metrics are shown in Fig. 1. For these simulated distances, we fixed $m = 100$ and let $p = 10, 100, 10000$ to see the convergence for different combinations of m and p . Normality was assessed using the Shapiro-Wilk test.

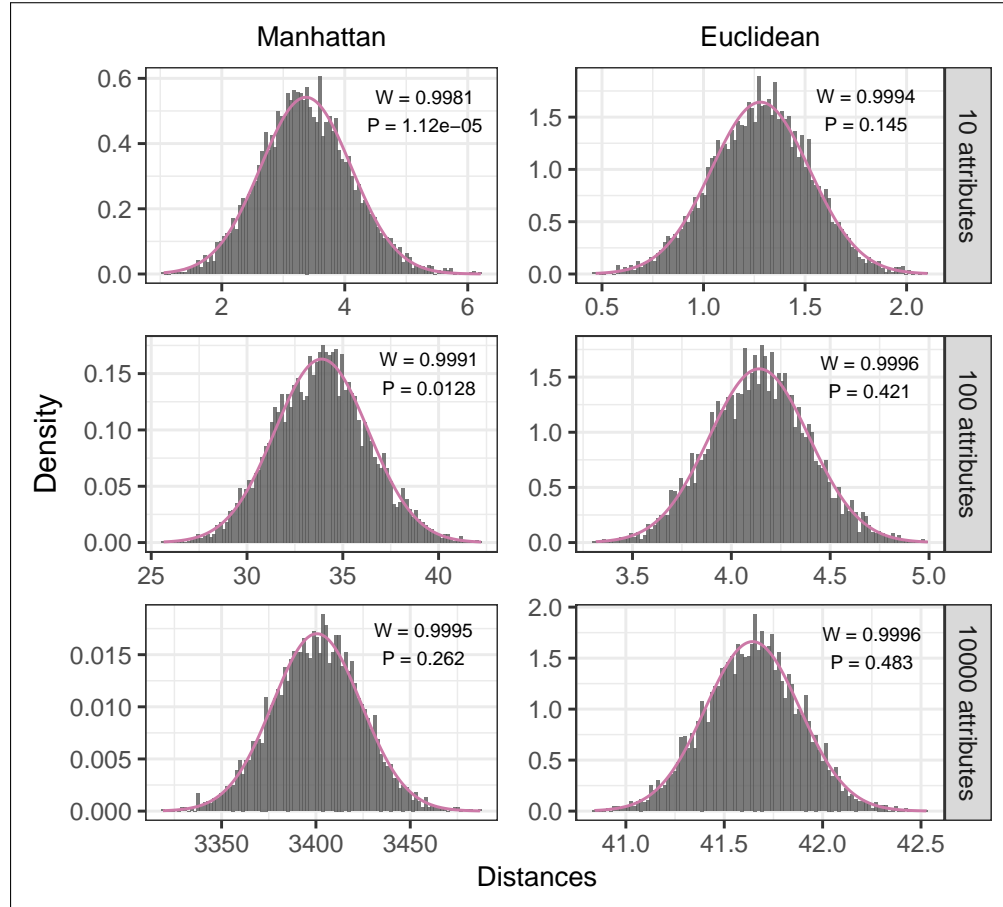


Fig 1. Convergence to normality of Manhattan and Euclidean distances. For each simulated distance distribution, we fixed $m = 100$ instances and let $p = 10, 100, 10000$. It is clear that convergence is rapid, and approximate normality can be safely assumed for even $p = 10$.

Although some p-values are significant at the 0.05 level for Manhattan ($q = 1$), a

visual inspection of the corresponding QQ-plots shown in Fig. 2 indicate the normality assumption holds reasonably well.

107
108

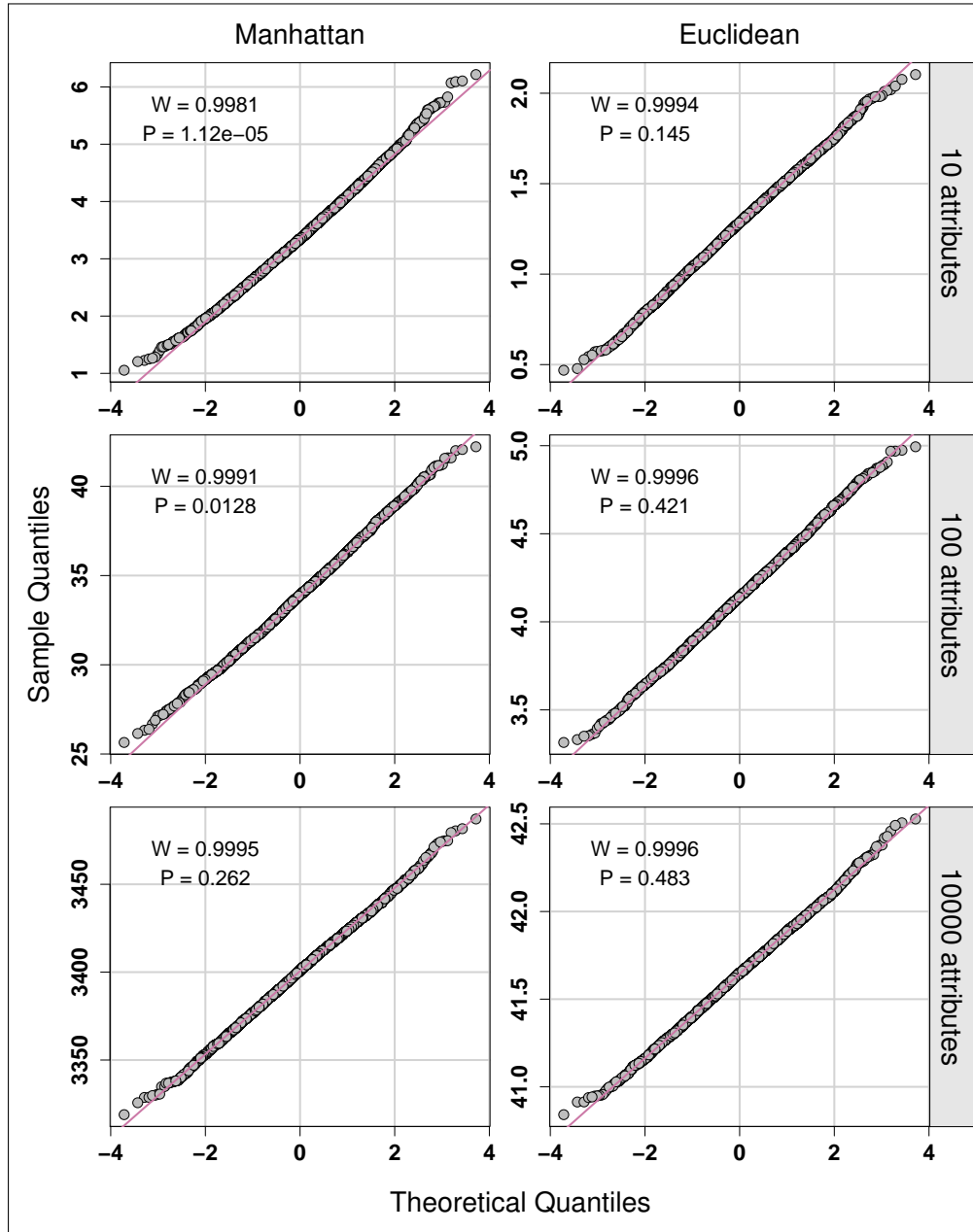


Fig 2. QQ-plots corresponding to the simulated distances in Fig. 1. Although there are significant p-values for the case of Manhattan for $p = 10, 100$, it is clear that the assumption of normality is safe due to strong relationship between sample and theoretical quantiles.

For distance based learning methods, all pairwise distances are used to determine relative importances for attributes. The collection of all distances above the diagonal in an $m \times m$ distance matrix does not satisfy the independence assumption used in the previous derivations. This is because of the redundancy that is inherent to the distance matrix calculation. However, this collection is still asymptotically normal with mean and

109
110
111
112
113

variance approximately equal to those given in Eq. 6. In the next section, we assume
 actual data distributions in order to define more specific general formulas for standard
 L_q and max-min normalized L_q metrics. We also derive asymptotic moments for a new
 discrete metric in GWAS data and a new metric for time series correlation-based data,
 such as, resting-state fMRI.

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

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

Suppose that $X_{ia}, X_{ja} \stackrel{iid}{\sim} \mathcal{F}_X(\mu_x, \sigma_x^2)$ 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 [11].

Theorem 2.1 *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*

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

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. 2.1 that the joint density function $g^{(1)}$ of X_{ia} and Z_a^q is given
 by

$$\begin{aligned} g^{(1)}(x_{ia}, z_a) &= f_X(x_{ia}, x_{ja}) \left| \frac{\partial x_{ja}}{\partial z_a} \right| \\ &= f_X(x_{ia}) f_X(x_{ja}) \left| \frac{-1}{q} (z_a^q)^{\frac{1}{q}-1} \right| \\ &= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} f_X(x_{ia}) f_X \left(x_{ia} - (z_a^q)^{1/q} \right), \quad z_a > 0 \end{aligned} \tag{8}$$

The density function $f_{Z_a^q}^{(1)}$ of Z_a^q is then defined as

$$\begin{aligned} f_{Z_a^q}^{(1)}(z_a^q) &= \int_{-\infty}^{\infty} g^{(1)}(x_{ia}, z_a^q) dx_{ia} \\ &= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) f_X \left(x_{ia} - (z_a^q)^{1/q} \right) dx_{ia}, \quad z_a > 0. \end{aligned} \tag{9}$$

- (ii) 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. 2.1 that the joint density function $g^{(2)}$ of X_{ia} and Z_a is given

by

$$\begin{aligned}
g^{(2)}(x_{ia}, z_a) &= f_X(x_{ia}, x_{ja}) \left| \frac{\partial x_{ja}}{\partial z_a} \right| \\
&= f_X(x_{ia}) f_X(x_{ja}) \left| \frac{1}{q} (z_a^q)^{\frac{1}{q}-1} \right| \\
&= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} f_X(x_{ia}) f_X(x_{ia} - (z_a^q)^{1/q}), \quad z_a > 0.
\end{aligned} \tag{10}$$

The density function $f_{Z_a^q}^{(2)}$ of Z_a^q is then defined as

$$\begin{aligned}
f_{Z_a^q}^{(2)}(z_a^q) &= \int_{-\infty}^{\infty} g^{(2)}(x_{ia}, z_a^q) dx_{ia} \\
&= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) f_X(x_{ia} + (z_a^q)^{1/q}) dx_{ia}, \quad z_a > 0.
\end{aligned} \tag{11}$$

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

$$E^{(1)} = \{|X_{ia} - X_{ja}|^q \leq z_a^q | X_{ja} = X_{ia} - (Z_a^q)^{1/q}\} \tag{12}$$

and

$$E^{(2)} = \{|X_{ia} - X_{ja}|^q \leq z_a^q | X_{ja} = X_{ia} + (Z_a^q)^{1/q}\}. \tag{13}$$

Then it follows from fundamental rules of probability that

$$\begin{aligned}
F_{Z_a^q}(z_a^q) &= \mathbb{P}[Z_a^q \leq z_a^q] \\
&= \mathbb{P}[|X_{ia} - X_{ja}|^q \leq z_a^q] \\
&= \mathbb{P}[E^{(1)} \cup E^{(2)}] \\
&= \mathbb{P}[E^{(1)}] + \mathbb{P}[E^{(2)}] - \mathbb{P}[E^{(1)} \cap E^{(2)}] \\
&= \mathbb{P}[E^{(1)}] + \mathbb{P}[E^{(2)}] \\
&= \int_{-\infty}^{z_a^q} f_{Z_a^q}^{(1)}(t) dt + \int_{-\infty}^{z_a^q} f_{Z_a^q}^{(2)}(t) dt \\
&= \int_{-\infty}^{z_a^q} (f_{Z_a^q}^{(1)}(t) + f_{Z_a^q}^{(2)}(t)) dt \\
&= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{z_a^q} \left(\int_{-\infty}^{\infty} f_X(x_{ia}) [f_X(x_{ia} - t) + f_X(x_{ia} + t)] dx_{ia} \right) dt, \quad z_a > 0.
\end{aligned} \tag{14}$$

It follows directly from the result in Eq. 14 that the density function of the random variable Z_a^q is given by

$$\begin{aligned}
f_{Z_a^q}(z_a^q) &= \frac{\partial}{\partial z_a^q} F_{Z_a^q}(z_a^q) \\
&= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[f_X(x_{ia} - (z_a^q)^{1/q}) + f_X(x_{ia} + (z_a^q)^{1/q}) \right] dx_{ia},
\end{aligned} \tag{15}$$

where $z_a > 0$.

Using Eq. 15, we can compute the mean and variance of the random variable Z_a^q as

$$\mu_{z^q} = \int_{-\infty}^{\infty} z_a^q f_{Z_a^q}(z_a^q) dz_a^q \tag{16}$$

and

$$\sigma_{z^q}^2 = \int_{-\infty}^{\infty} (z_a^q)^2 f_{Z^q}(z_a^q) dz_a^q - \mu_{z^q}^2. \quad (17)$$

It follows immediately from Eqs. 16 and 17 and the Classical Central Limit Theorem (CCLT) that

$$\left(D_{ij}^{(q)}\right)^q = \sum_{a \in \mathcal{A}} Z_a^q = \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \sim \mathcal{N}(\mu_{z^q} p, \sigma_{z^q}^2 p). \quad (18)$$

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

$$D_{ij}^{(q)} \sim \mathcal{N}\left((\mu_{z^q} p)^{1/q}, \frac{\sigma_{z^q}^2 p}{q^2 (\mu_{z^q} p)^{2(1-\frac{1}{q})}}\right), \quad \mu_{z^q} > 0 \quad (19)$$

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

2.1.1 Standard normal data

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

$$f_X(x_{ia}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_{ia}^2}, \quad (20)$$

$$f_X\left(x_{ia} - (z_a^q)^{1/q}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_{ia} - (z_a^q)^{1/q})^2}, \quad z_a > 0, \text{ and} \quad (21)$$

$$f_X\left(x_{ia} + (z_a^q)^{1/q}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_{ia} + (z_a^q)^{1/q})^2}, \quad z_a > 0. \quad (22)$$

Substituting the results given by Eqs. 20-22 into Eq. 15 and completing the square on x_{ia} in the exponents, we have

$$f_{Z^q}(z_a^q) = \frac{1}{2q\pi (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \left(e^{-\frac{1}{2}[\sqrt{2}x_{ia} - \frac{\sqrt{2}}{2}(z_a^q)^{1/q}]^2} + e^{-\frac{1}{2}[\sqrt{2}x_{ia} + \frac{\sqrt{2}}{2}(z_a^q)^{1/q}]^2} \right) dx_{ia} \quad (23)$$

$$= \frac{1}{2q\sqrt{\pi} (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left(e^{-\frac{1}{2}u^2} + e^{-\frac{1}{2}u^2} \right) du \quad (24)$$

$$= \frac{1}{2q\sqrt{\pi} (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} (1 + 1) \quad (25)$$

$$= \frac{1}{q\sqrt{\pi}} (z_a^q)^{\frac{1}{q}-1} e^{-\frac{1}{4}(z_a^q)^{2/q}} \quad (26)$$

$$= \frac{\frac{2}{q}}{(2q)^{1/q} \Gamma\left(\frac{1}{q}\right)} (z_a^q)^{\frac{1}{q}-1} e^{-\left(\frac{z_a^q}{2q}\right)^{2/q}}. \quad (27)$$

The density function given by Eq. 23 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

$$\begin{aligned} \mu_{z^q} &= \frac{c\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \\ &= \frac{2^q \Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} \end{aligned} \quad (28)$$

and

$$\begin{aligned}\sigma_{z^q}^2 &= c^2 \left[\frac{\Gamma\left(\frac{d+2}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} - \left(\frac{\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \right)^2 \right] \\ &= 4^q \left[\frac{\Gamma\left(q + \frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q + \frac{1}{2}\right)}{\pi} \right].\end{aligned}\quad (29)$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 28 and 29 allow the p -dimensional mean and variance of the $D_{ij}^{(q)}$ distribution to be computed directly as

$$\mu_{(D_{ij}^{(q)})^q} = \mathbb{E} \left[(D_{ij}^{(q)})^q \right] = \mathbb{E} \left(\sum_{a \in \mathcal{A}} Z_a^q \right) = \sum_{a \in \mathcal{A}} \mathbb{E}(Z_a^q) = \sum_{a \in \mathcal{A}} \frac{2^q \Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} = \frac{2^q \Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} p \quad (30)$$

and

$$\begin{aligned}\sigma_{(D_{ij}^{(q)})^q}^2 &= \text{Var} \left[(D_{ij}^{(q)})^q \right] = \text{Var} \left(\sum_{a \in \mathcal{A}} Z_a^q \right) \\ &= \sum_{a \in \mathcal{A}} \text{Var}(Z_a^q) \\ &= \sum_{a \in \mathcal{A}} 4^q \left[\frac{\Gamma\left(q + \frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q + \frac{1}{2}\right)}{\pi} \right] \\ &= 4^q \left[\frac{\Gamma\left(q + \frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q + \frac{1}{2}\right)}{\pi} \right] p.\end{aligned}\quad (31)$$

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

$$\mathcal{N} \left(\left(2^q \frac{\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} p \right)^{1/q}, \frac{4^q p}{q^2 \left(\frac{2^q \Gamma\left(\frac{1}{2}q + \frac{1}{2}\right)}{\sqrt{\pi}} p \right)^{2(1-\frac{1}{q})}} \left[\frac{\Gamma\left(q + \frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q + \frac{1}{2}\right)}{\pi} \right] \right). \quad (32)$$

2.1.2 Standard uniform data

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

$$f_X(x_{ia}) = 1, \quad 0 \leq x_{ia} \leq 1 \quad (33)$$

$$f_X \left(x_{ia} - (z_a^q)^{1/q} \right) = 1, \quad 0 \leq x_{ia} - (z_a^q)^{1/q} \leq 1, \text{ and} \quad (34)$$

$$f_X \left(x_{ia} + (z_a^q)^{1/q} \right) = 1, \quad 0 \leq x_{ia} + (z_a^q)^{1/q} \leq 1. \quad (35)$$

Substituting the results given by Eqs. 33-35 into Eq. 15, we have

172

$$\begin{aligned}
f_{Z^q}(z_a^q) &= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[f_X(x_{ia} - (z_a^q)^{1/q}) + f_X(x_{ia} + (z_a^q)^{1/q}) \right] dx_{ia}, \\
&\quad 0 < z_a \leq 1 \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_0^1 [f_X(x_{ia} - (z_a^q)^{1/q}) + f_X(x_{ia} + (z_a^q)^{1/q})] dx_{ia}, \quad 0 < z_a \leq 1 \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_{(z_a^q)}^1 1 dx_{ia} + \int_0^{1-(z_a^q)} 1 dx_{ia}, \quad 0 < z_a \leq 1 \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} [(1 - (z_a^q)) + (1 - (z_a^q))], \quad 0 < z_a \leq 1 \\
&= \frac{1}{q} \cdot 2(z_a^q)^{\frac{1}{q}-1} [1 - (z_a^q)^{1/q}]^{2-1}, \quad 0 < z_a \leq 1.
\end{aligned} \tag{36}$$

The density given by Eq. 36 is a Kumaraswamy density with parameters $b = \frac{1}{q}$ and $c = 2$ with moment generating function (MGF) given by

173

174

$$\begin{aligned}
M_n &= \frac{c\Gamma(1 + \frac{n}{b})\Gamma(c)}{\Gamma(1 + c + \frac{n}{b})} \\
&= \frac{2}{(nq + 2)(nq + 1)}.
\end{aligned} \tag{37}$$

Using the MGF given by Eq. 37, the mean and variance of Z_a^q are computed as

175

$$\mu_{z^q} = \frac{2}{(q + 2)(q + 1)} \tag{38}$$

and

176

$$\sigma_{z^q}^2 = \frac{1}{(q + 1)(2q + 1)} - \left(\frac{2}{(q + 2)(q + 1)} \right)^2. \tag{39}$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 40 and 41 allow the p -dimensional mean and variance of the $(D_{ij}^{(q)})^q$ distribution to be computed directly as

177

178

179

$$\begin{aligned}
\mu_{(D_{ij}^{(q)})^q} &= E \left[(D_{ij}^{(q)})^q \right] = E \left(\sum_{a \in \mathcal{A}} Z_a^q \right) \\
&= \sum_{a \in \mathcal{A}} E(Z_a^q) \\
&= \sum_{a \in \mathcal{A}} \frac{2}{(q + 2)(q + 1)} \\
&= \frac{2p}{(q + 2)(q + 1)}
\end{aligned} \tag{40}$$

and

180

$$\begin{aligned}
\sigma^2_{(D_{ij}^{(q)})^q} &= \text{Var} \left[\left(D_{ij}^{(q)} \right)^q \right] = \text{Var} \left(\sum_{a \in \mathcal{A}} Z_a^q \right) \\
&= \sum_{a \in \mathcal{A}} \text{Var} (Z_a^q) \\
&= \sum_{a \in \mathcal{A}} \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)} \right)^2 \right] \\
&= \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)} \right)^2 \right] p.
\end{aligned} \tag{41}$$

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

181

$$\begin{aligned}
&\mathcal{N} \left(\left(\frac{2p}{(q+2)(q+1)} \right)^{1/q}, \right. \\
&\quad \left. \frac{p}{q^2 \left(\frac{2p}{(q+2)(q+1)} \right)^{2(1-\frac{1}{q})}} \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)} \right)^2 \right] \right).
\end{aligned} \tag{42}$$

2.2 Manhattan ($q = 1$)

182

With our general formulas for the asymptotic mean and variance given by Eqs. 32 and 42 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.

183

184

185

186

187

2.2.1 Standard normal data

188

Using the mean given by Eq. 32 and substituting $q = 1$, we have the following for standard normal data

189

190

$$\begin{aligned}
\mathbb{E} \left(D_{ij}^{(1)} \right) &= \left(2 \frac{\Gamma \left(\frac{1+1}{2} \right)}{\sqrt{\pi}} p \right)^{1/1} \\
&= \frac{2p}{\sqrt{\pi}} \Gamma(1) \\
&= \frac{2p}{\sqrt{\pi}}.
\end{aligned} \tag{43}$$

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

191

$$\begin{aligned}
\text{Var} \left(D_{ij}^{(1)} \right) &= \frac{4^1 p}{1^2 \left(\frac{2^1 \Gamma(\frac{1}{2}(1) + \frac{1}{2})}{\sqrt{\pi}} p \right)^{2(1-\frac{1}{1})}} \left[\frac{\Gamma(1 + \frac{1}{2})}{\sqrt{\pi}} - \frac{\Gamma^2(\frac{1}{2}(1) + \frac{1}{2})}{\pi} \right] \\
&= \frac{4p}{1} \left[\frac{\frac{1}{2} \Gamma(\frac{1}{2})}{\sqrt{\pi}} - \frac{\Gamma^2(1)}{\pi} \right] \\
&= 4p \left[\frac{1}{2} - \frac{1}{\pi} \right] \\
&= \frac{2(\pi - 2)p}{\pi}.
\end{aligned} \tag{44}$$

2.2.2 Standard uniform data

192

Using the mean given by Eq. 42 and substituting $q = 1$, we have the following for standard uniform data

193

194

$$\begin{aligned}
\text{E} \left(D_{ij}^{(1)} \right) &= \left(\frac{2p}{(1+2)(1+1)} \right)^{1/1} \\
&= \frac{2p}{6} \\
&= \frac{p}{3}.
\end{aligned} \tag{45}$$

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

195

$$\begin{aligned}
\text{Var} \left(D_{ij}^{(1)} \right) &= \frac{p}{1^2 \left(\frac{2p}{(1+2)(1+1)} \right)^{2(1-\frac{1}{1})}} \left[\frac{1}{(1+1)(2(1)+1)} - \left(\frac{2}{(1+2)(1+1)} \right)^2 \right] \\
&= p \left[\frac{1}{6} - \frac{1}{9} \right] \\
&= \frac{p}{18}.
\end{aligned} \tag{46}$$

2.3 Euclidean ($q = 2$)

196

Analogous to the previous section, we demonstrate the usage of Eqs. 32 and 42 for the Euclidean ($q = 2$) metric for standard normal and standard uniform data.

197

198

2.3.1 Standard normal data

199

Using the mean given by Eq. 32 and substituting $q = 2$, we have the following for standard normal data

200

201

$$\begin{aligned}
\text{E} \left(D_{ij}^{(2)} \right) &= \left(2 \frac{\Gamma(\frac{2+1}{2})}{\sqrt{\pi}} p \right)^{1/2} \\
&= \left(\frac{2p}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) \right)^{1/2} \\
&= \sqrt{2p}.
\end{aligned} \tag{47}$$

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

202

$$\begin{aligned}
\text{Var} \left(D_{ij}^{(1)} \right) &= \frac{4^2 p}{2^2 \left(\frac{2^2 \Gamma(\frac{1}{2}(2) + \frac{1}{2})}{\sqrt{\pi}} p \right)^{2(1-\frac{1}{2})}} \left[\frac{\Gamma(2 + \frac{1}{2})}{\sqrt{\pi}} - \frac{\Gamma^2(\frac{1}{2}(2) + \frac{1}{2})}{\pi} \right] \\
&= \frac{16p}{4 \left(\frac{4\Gamma(\frac{3}{2})}{\sqrt{\pi}} p \right)} \left[\frac{\Gamma(\frac{5}{2})}{\sqrt{\pi}} - \frac{\Gamma^2(\frac{3}{2})}{\pi} \right] \\
&= 2 \left[\frac{3}{4} - \frac{1}{4} \right] \\
&= 1.
\end{aligned} \tag{48}$$

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

203

204

$$\begin{aligned}
\text{E} \left(D_{ij}^{(2)} \right) &= \left(2 \frac{\Gamma(\frac{2+1}{2})}{\sqrt{\pi}} p - 1 \right)^{1/2} \\
&= \left(\frac{2p}{\sqrt{\pi}} \Gamma\left(\frac{3}{2}\right) - 1 \right)^{1/2} \\
&= \sqrt{2p - 1}.
\end{aligned} \tag{49}$$

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. 47 since $\sqrt{2p} \approx \sqrt{2p - 1}$ in that case.

205

206

207

2.3.2 Standard uniform data

208

Using the mean given by Eq. 42 and substituting $q = 2$, we have the following for standard uniform data

209

210

$$\begin{aligned}
\text{E} \left(D_{ij}^{(2)} \right) &= \left(\frac{2p}{(2+2)(2+1)} \right)^{1/2} \\
&= \left(\frac{2p}{12} \right)^{1/2} \\
&= \sqrt{\frac{p}{6}}.
\end{aligned} \tag{50}$$

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

211

$$\begin{aligned}
\text{Var} \left(D_{ij}^{(2)} \right) &= \frac{p}{2^2 \left(\frac{2p}{(2+2)(2+1)} \right)^{2(1-\frac{1}{2})}} \left[\frac{1}{(2+1)(2(2)+1)} - \left(\frac{2}{(2+2)(2+1)} \right)^2 \right] \\
&= \frac{3}{2} \left[\frac{1}{15} - \frac{1}{36} \right] \\
&= \frac{7}{120}.
\end{aligned} \tag{51}$$

For the case in which the number of attributes p is small, an improved estimate of

212

the mean is given by Eq. 7. The lower dimensional estimate of the mean is as follows 213

$$\begin{aligned} \mathbb{E} \left(D_{ij}^{(2)} \right) &= \left(\frac{2p}{(2+2)(2+1)} - \frac{7}{120} \right)^{1/2} \\ &= \left(\frac{2p}{12} - \frac{7}{120} \right)^{1/2} \\ &= \sqrt{\frac{p}{6} - \frac{7}{120}}. \end{aligned} \tag{52}$$

For high dimensional data sets, such as gene expression, rs-fMRI, or GWAS, it is 214
clear that the magntiude of p will be sufficient to use Eq. 47 since $\sqrt{\frac{p}{6}} \approx \sqrt{\frac{p}{6} - \frac{7}{120}}$ in 215
that case. 216

2.4 Distribution of attribute extremes 217

For Relief-based methods [3, 12], the standard numeric diff metric is given by 218

$$d_{ij}^{\text{num}}(a) = \text{diff}(a, (i, j)) = \frac{|X_{ia} - X_{ja}|}{\max(a) - \min(a)}, \tag{53}$$

where $\max(a) = \max_{k \in \mathcal{I}} \{X_{ka}\}$, $\min(a) = \min_{k \in \mathcal{I}} \{X_{ka}\}$, and $\mathcal{I} = \{1, 2, \dots, m\}$. 219

In order to determine moments of asymptotic distance distributions induced by Eq. 53, 220
we must first derive the asymptotic extreme value distributions of the attribute maximum 221
and minimum. Although the exact distribution of the maximum or minimum requires an 222
assumption about the data distribution, the Fisher-Tippett-Gnedenko Theorem allows 223
us to categorize the extreme value distribution for a collection of independent and 224
identically distributed random variables into one of three distributional families. Before 225
stating the theorem, we first need the following definition. 226

Definition 2.1 A distribution \mathcal{F}_X is said to be **degenerate** if its density function f_X 227
is the Dirac delta $\delta(x - c_0)$ centered at a constant $c_0 \in \mathbb{R}$, with corresponding distribution 228
function F_X defined as 229

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

Theorem 2.2 (Fisher-Tippett-Gnedenko) Let $X_{1a}, X_{2a}, \dots, X_{ma} \stackrel{iid}{\sim} \mathcal{F}_X(\mu_x, \sigma_x^2)$ 230
and let $X_a^{\max} = \max_{k \in \mathcal{I}} \{X_{ka}\}$. If there exists two non-random sequences $b_m > 0$ and c_m 231
such that 232

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

where G_X is a non-degenerate distribution function, then the limiting distribution \mathcal{G}_X is 233
in the Gumbel, Fréchet, or Weibull family. 234

The three distribution families given in Thm. 2.2 are actually special cases of the 235
Generalized Extreme Value Distribution. In the context of extreme values, Thm. 2.2 is 236
analogous to the Central Limit Theorem for the distribution of sample mean. We will 237
take advantage of this theorem for the distribution of the maximum for standard normal 238
data to show that the limiting distribution is in the Gumbel family. However, we will 239

derive the distribution of the maximum and minimum for standard uniform data directly. 240
 Regardless of data type, the distribution of the sample maximum is derived as follows 241

$$\begin{aligned}
 P[X_a^{\max} \leq x] &= P\left[\max_{k \in \mathcal{I}}\{X_{ka}\} \leq x\right] \\
 &= P[X_{1a} \leq x, X_{2a} \leq x, \dots, X_{ma} \leq x] \\
 &= \prod_{k=1}^m P[X_{ka} \leq x] \\
 &= \prod_{k=1}^m F_X(x) \\
 &= [F_X(x)]^m.
 \end{aligned} \tag{54}$$

Therefore, we have the following expression for the distribution function of the 242
 maximum 243

$$F_{\max}(x) = [F_X(x)]^m. \tag{55}$$

Differentiating the distribution function given by Eq. 55 gives us the following density 244
 function for the distribution of the maximum 245

$$\begin{aligned}
 f_{\max}(x) &= \frac{d}{dx} F_{\max}(x) \\
 &= \frac{d}{dx} [F_X(x)]^m \\
 &= m[F_X(x)]^{m-1} f_X(x).
 \end{aligned} \tag{56}$$

The distribution of the sample minimum, X_a^{\min} , is derived as follows 246

$$\begin{aligned}
 P[X_a^{\min} \leq x] &= 1 - P[X_a^{\min} \geq x] \\
 &= 1 - P\left[\min_{k \in \mathcal{I}}\{X_{ka}\} \geq x\right] \\
 &= 1 - P[X_{1a} \geq x, X_{2a} \geq x, \dots, X_{ma} \geq x] \\
 &= 1 - \prod_{k=1}^m P[X_{ka} \geq x] \\
 &= 1 - [P[X_{1a} \geq x]]^m \\
 &= 1 - [1 - P[X_{1a} \leq x]]^m \\
 &= 1 - [1 - F_X(x)]^m.
 \end{aligned} \tag{57}$$

Therefore, we have the following expression for the distribution function of the 247
 minimum 248

$$F_{\min}(x) = 1 - [1 - F_X(x)]^m. \tag{58}$$

Differentiating the distribution function given by Eq. 58 gives us the following density 249
 function for the distribution of the minimum 250

$$\begin{aligned}
 f_{\min}(x) &= \frac{d}{dx} F_{\min}(x) \\
 &= \frac{d}{dx} (1 - [1 - F_X(x)]^m) \\
 &= m[1 - F_X(x)]^{m-1} f_X(x).
 \end{aligned} \tag{59}$$

Given the densities of the distribution of sample maximum and minimum, we can 251
 easily compute moments and the variance. The first and second moment about the 252

origin and the variance of the distribution of the maximum are given by the following 253

$$\begin{aligned}\mu_{\max}^{(1)}(m) &= E(X_a^{\max}) = \int_{-\infty}^{\infty} x f_{\max}(x) dx \\ &= \int_{-\infty}^{\infty} x (m[F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x f_X(x) [F_X(x)]^{m-1} dx.\end{aligned}\tag{60}$$

254

$$\begin{aligned}\mu_{\max}^{(2)}(m) &= E[(X_a^{\max})^2] = \int_{-\infty}^{\infty} x^2 f_{\max}(x) dx \\ &= \int_{-\infty}^{\infty} x^2 (m[F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x^2 f_X(x) [F_X(x)]^{m-1} dx\end{aligned}\tag{61}$$

255

$$\sigma_{\max}^2(m) = \mu_{\max}^{(2)}(m) - [\mu_{\max}^{(1)}(m)]^2\tag{62}$$

Similarly, we have the first and second moment about the origin and variance of the distribution of sample minimum given by the following 256
257

$$\begin{aligned}\mu_{\min}^{(1)}(m) &= E(X_a^{\min}) = \int_{-\infty}^{\infty} x f_{\min}(x) dx \\ &= \int_{-\infty}^{\infty} x (m[F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x f_X(x) [F_X(x)]^{m-1} dx,\end{aligned}\tag{63}$$

258

$$\begin{aligned}\mu_{\min}^{(2)}(m) &= E[(X_a^{\min})^2] = \int_{-\infty}^{\infty} x^2 f_{\min}(x) dx \\ &= \int_{-\infty}^{\infty} x^2 (m[F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x^2 f_X(x) [F_X(x)]^{m-1} dx,\end{aligned}\tag{64}$$

and

$$\sigma_{\min}^2(m) = \mu_{\min}^{(2)}(m) - [\mu_{\min}^{(1)}(m)]^2.\tag{65}$$

259

With the densities of attribute maximum and minimum for sample size m , the expected range is given by the following 260
261

$$\begin{aligned}E(X_a^{\max} - X_a^{\min}) &= E(X_a^{\max}) - E(X_a^{\min}) \\ &= \mu_{\max}^{(1)}(m) - \mu_{\min}^{(1)}(m).\end{aligned}\tag{66}$$

For a data distribution that has zero skewness and has support that is symmetric about 0, the result given by Eq. 66 can be simplified to the following expression 262
263

$$E(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m).\tag{67}$$

For large samples ($m \gg 1$), the covariance between the sample maximum and minimum is approximately zero [13]. Therefore, the variance of the attribute range of a sample of size m is given by the following 264
265
266

$$\begin{aligned}\text{Var}(X_a^{\max} - X_a^{\min}) &\approx \text{Var}(X_a^{\max}) + \text{Var}(X_a^{\min}) \\ &= \sigma_{\max}^2(m) + \sigma_{\min}^2(m).\end{aligned}\tag{68}$$

Under the assumption of zero skewness and support that is symmetric about 0, the result given by Eq. 68 becomes the following

$$\begin{aligned}\text{Var}(X_a^{\max} - X_a^{\min}) &= 2\text{Var}(X_a^{\max}) \\ &= 2\sigma_{\max}^2.\end{aligned}\tag{69}$$

Let $\mu_{D_{ij}^{(q)}}$ and $\sigma_{D_{ij}^{(q)}}^2$ denote the mean and variance given by Eq. 19. 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. 53. Then the mean of the max-min normalized distance distribution is given by the following

$$\begin{aligned}\mu_{D_{ij}^{(q*)}} &= \mathbb{E} \left[\left(\sum_{a \in \mathcal{A}} \left(\frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{1/q} \right] \\ &\approx \frac{1}{\mathbb{E}(X_a^{\max} - X_a^{\min})} \mathbb{E} \left[\left(\sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \right)^{1/q} \right] \\ &= \frac{\mu_{D_{ij}^{(q)}}}{\mathbb{E}(X_a^{\max}) - \mathbb{E}(X_a^{\min})} \\ &= \frac{\mu_{D_{ij}^{(q)}}}{\mu_{\max}^{(1)} - \mu_{\min}^{(1)}}.\end{aligned}\tag{70}$$

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

$$\begin{aligned}\sigma_{D_{ij}^{(q*)}}^2 &= \text{Var} \left[\left(\sum_{a \in \mathcal{A}} \left(\frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{1/q} \right] \\ &= \mathbb{E} \left[\left(\sum_{a \in \mathcal{A}} \left(\frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{2/q} \right] - \left(\mathbb{E} \left[\left(\sum_{a \in \mathcal{A}} \left(\frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{1/q} \right] \right)^2 \\ &\approx \frac{\mathbb{E} \left[\left(\sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \right)^{2/q} \right]}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} - \left(\frac{\mathbb{E} \left[\left(\sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \right)^{1/q} \right]}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} \right)^2 \\ &= \frac{\sigma_{D_{ij}^{(q)}}^2 + \mu_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} - \frac{\mu_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} \\ &= \frac{\sigma_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} \\ &= \frac{\sigma_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max})^2] - 2\mathbb{E}(X_a^{\max})\mathbb{E}(X_a^{\min}) + \mathbb{E}(X_a^{\min})^2}} \\ &= \frac{\sigma_{D_{ij}^{(q)}}^2}{\mu_{\max}^{(2)}(m) - 2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m) + \mu_{\min}^{(2)}(m)}.\end{aligned}\tag{71}$$

With the results given by Eqs. 70 and 71, we have the following generalized estimate for the asymptotic distribution of the max-min normalized distance distribution

$$D_{ij}^{(q*)} \sim \mathcal{N} \left(\frac{\mu_{D_{ij}^{(q)}}}{\mu_{\max}^{(1)}(m) - \mu_{\min}^{(1)}(m)}, \frac{\sigma_{D_{ij}^{(q)}}^2}{\mu_{\max}^{(2)}(m) - 2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m) + \mu_{\min}^{(2)}(m)} \right).\tag{72}$$

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. 70 exclusively in terms of the expected maximum. This result is given by the following

$$\mu_{D_{ij}^{(q*)}} \approx \frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}. \quad (73)$$

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

$$\begin{aligned} \sigma_{D_{ij}^{(q*)}}^2 &\approx \frac{\sigma_{D_{ij}^{(q)}}^2}{2\mu_{\max}^{(2)}(m) + 2\left[\mu_{\max}^{(1)}(m)\right]^2} \\ &= \frac{\sigma_{D_{ij}^{(q)}}^2}{2\left(\sigma_{\max}^2(m) + \left[\mu_{\max}^{(1)}(m)\right]^2\right)}. \end{aligned} \quad (74)$$

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

$$D_{ij}^{(q*)} \sim \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}, \frac{\sigma_{D_{ij}^{(q)}}^2}{2\left(\sigma_{\max}^2(m) + \left[\mu_{\max}^{(1)}(m)\right]^2\right)}\right). \quad (75)$$

2.4.1 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. 60 and 61 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

$$\begin{aligned} \log\Phi(-c_m - b_mx) &= \log\Phi(-c_m) - b_mx \frac{\phi(-c_m)}{\Phi(-c_m)} + \mathcal{O}(b_m^2 x^2) \\ &= \log\left(\frac{1}{m}\right) - x \frac{\phi(-c_m)}{c_m \Phi(-c_m)} + \mathcal{O}(b_m^2 x^2). \end{aligned} \quad (76)$$

In order to simplify the right-hand side of Eq. 76, we will use the well known Mills Ratio Bounds [14] given by the following

$$1 \leq \frac{\phi(x)}{x\Phi(-x)} \leq 1 + \frac{1}{x^2}, \quad x > 0. \quad (77)$$

The inequalities given by Eq. 77 show that $\frac{\phi(x)}{x\Phi(-x)} \rightarrow 1$ as $x \rightarrow \infty$. This implies that $\frac{\phi(c_m)}{c_m \Phi(-c_m)} \rightarrow 1$ as $m \rightarrow \infty$ since $c_m = -\Phi^{-1}\left(\frac{1}{m}\right) \rightarrow \infty$ as $m \rightarrow \infty$. This gives us the

following approximation of the right-hand side of Eq. 76

299

$$\begin{aligned}
\log \Phi(-c_m - b_m x) &\approx \log \left(\frac{1}{m} \right) - x + \mathcal{O}(b_m^2 x^2) \\
\Rightarrow \Phi(-c_m - b_m x) &\approx \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)} \\
\Rightarrow \Phi(c_m + b_m x) &\approx 1 - \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)}.
\end{aligned} \tag{78}$$

Using the result given by Eq. 78, we have the following

300

$$\begin{aligned}
P \left(\frac{X_a^{\max} - c_m}{b_m} \leq x \right) &= P(X_a^{\max} \leq c_m + b_m x) \\
&= \Phi^m(c_m + b_m x) \\
&\approx \left(1 - \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)} \right)^m \\
&= \left(1 - \frac{1}{m} e^{-x + \mathcal{O}(\frac{1}{m} x^2)} \right)^m \\
&\approx \left(1 - \frac{1}{m} e^{-x} \right)^m \\
\Rightarrow \lim_{m \rightarrow \infty} P \left(\frac{X_a^{\max} - c_m}{b_m} \leq x \right) &= \lim_{m \rightarrow \infty} \left(1 - \frac{1}{m} e^{-x} \right)^m \\
&= e^{-e^{-x}}.
\end{aligned} \tag{79}$$

The right-hand side of Eq. 79 is the cumulative distribution function of the standard Gumbel distribution. The mean of the asymptotic distribution is given by the following

301

302

$$E(X_a^{\max}) = \mu_{\max}^{(1)} = -\Phi^{-1} \left(\frac{1}{m} \right) - \frac{\gamma}{\Phi^{-1} \left(\frac{1}{m} \right)}. \tag{80}$$

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

303

304

$$\tilde{\mu}_{\max} = \frac{\log(\log(2))}{\Phi^{-1} \left(\frac{1}{m} \right)} - \Phi^{-1} \left(\frac{1}{m} \right). \tag{81}$$

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

305

306

$$\text{Var}(X_a^{\max}) = \frac{\pi^2}{6} \left(\frac{1}{-\Phi^{-1} \left(\frac{1}{m} \right)} \right)^2. \tag{82}$$

For typical sample sizes m in high-dimensional spaces, the variance estimate given by Eq. 82 exceeds the variance of the sample maximum significantly. Using the fact that $-\Phi^{-1} \left(\frac{1}{m} \right) \sim \sqrt{2 \log(m)}$ [15] 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

307

308

309

310

$$\begin{aligned}
\sigma_{\max}^2(m) = \text{Var}(X_a^{\max}) &\approx \frac{\pi^2}{6} \left(\frac{1}{\sqrt{2 \log(m)}} \right)^2 \\
&= \frac{\pi^2}{12 \log(m)}.
\end{aligned} \tag{83}$$

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

$$E(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m) = 2 \left[-\Phi^{-1} \left(\frac{1}{m} \right) - \frac{\gamma}{\Phi^{-1} \left(\frac{1}{m} \right)} \right]. \quad (84)$$

It is well known that the sample extremes from the standard normal distribution
are approximately uncorrelated for large sample size m [13]. This implies that we can
approximate the variance of the range of m iid standard normal random variables with
the following result

$$\begin{aligned} \text{Var}(X_a^{\max} - X_a^{\min}) &\approx \text{Var}(X_a^{\max}) + \text{Var}(X_a^{\min}) \\ &= \sigma_{\max}^2(m) + \sigma_{\min}^2(m) \\ &= 2\sigma_{\max}^2(m) \\ &\approx 2 \left(\frac{\pi^2}{2\log(m)} \right) \\ &= \frac{\pi^2}{6\log(m)}. \end{aligned} \quad (85)$$

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. 80 overestimates the sample maximum. The formula for the
median of the sample maximum, given by Eq. 81, provides a more accurate estimate of
this sample extreme. Therefore, the following estimate for the mean of the attribute
range will be used instead

$$E(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m) \approx 2 \left[\frac{\log(\log(2))}{\Phi^{-1} \left(\frac{1}{m} \right)} - \Phi^{-1} \left(\frac{1}{m} \right) \right]. \quad (86)$$

We have already determined that $\mu_{D_{ij}}^{(q)}$ and $\sigma_{D_{ij}}^2$ are given by Eq. 32. Using the
results given by Eqs. 86 and 85 and the general formulas for the mean and variance
of the max-min normalized distance distribution given in Eq. 75, this leads us to the
following asymptotic estimate for the distribution of the max-min normalized distances
for standard normal data

$$D_{ij}^{(q*)} \sim \mathcal{N} \left(\frac{\mu_{D_{ij}}^{(q)}}{2\mu_{\max}^{(1)}(m)}, \frac{6\log(m)\sigma_{D_{ij}}^2}{\pi^2 + 24 \left[\mu_{\max}^{(1)}(m) \right]^2 \log(m)} \right). \quad (87)$$

2.4.2 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. 55 and 58, we have
the following distribution functions for standard uniform data

$$F_{\max}(x) = x^m \quad (88)$$

and

$$F_{\min}(x) = 1 - (1 - x)^m. \quad (89)$$

Using the general forms of the density functions of the maximum and minimum given by Eqs. 56 and 59, we have the following density functions for standard uniform data

$$f_{\max}(x) = mx^{m-1} \quad (90)$$

and

$$f_{\min}(x) = m(1-x)^{m-1} \quad (91)$$

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

$$\begin{aligned} E(X_a^{\max}) &= \mu_{\max}^{(1)}(m) = \int_0^1 x f_{\max}(x) dx \\ &= \int_0^1 x [mx^{m-1}] dx \\ &= \frac{m}{m+1} \end{aligned} \quad (92)$$

and

$$\begin{aligned} E(X_a^{\min}) &= \mu_{\min}^{(1)}(m) = \int_0^1 x f_{\min}(x) dx \\ &= \int_0^1 x [m(1-x)^{m-1}] dx \\ &= \frac{1}{m+1}. \end{aligned} \quad (93)$$

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

$$\begin{aligned} E[(X_a^{\max} - X_a^{\min})^2] &= E[(X_a^{\max})^2 - 2X_a^{\max}X_a^{\min} + (X_a^{\min})^2] \\ &= E[(X_a^{\max})^2] - 2E(X_a^{\max})E(X_a^{\min}) + E[(X_a^{\min})^2] \\ &= \mu_{\max}^{(2)}(m) - 2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m) + \mu_{\min}^{(2)}(m) \\ &= \int_0^1 x^2 [mx^{m-1}] dx - 2 \left(\frac{m}{m+1} \right) \left(\frac{1}{m+1} \right) \\ &\quad + \int_0^1 x^2 [m(1-x)^{m-1}] dx \\ &= \frac{m}{m+2} - \frac{2m}{(m+1)^2} + \frac{2}{(m+1)(m+2)} \\ &= \frac{m^3 - m + 2}{(m+2)(m+1)^2}. \end{aligned} \quad (94)$$

Using the general formulas given in Eq. 72 and the mean ($\mu_{D_{ij}^{(q)}}$) and variance ($\sigma_{D_{ij}^{(q)}}^2$) given by Eq. 42, we have the following asymptotic estimate for the max-min normalized distance distribution for standard uniform data

$$D_{ij}^{(q*)} \sim \mathcal{N} \left(\frac{(m+1)\mu_{D_{ij}^{(q)}}}{m-1}, \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(q)}}^2}{m^3 - m + 2} \right). \quad (95)$$

2.5 Normalized Manhattan ($q = 1$)

Using the general asymptotic results for mean and variance given by Eqs. 87 and 95 for any value of $q \in \mathbb{N}$, we can substitute a particular value of q in order to determine a more specified asymptotic distance distribution for $D_{ij}^{(q*)}$. The following results are for the max-min normalized Manhattan ($q = 1$) metric for both standard normal and standard uniform data.

2.5.1 Standard normal data

Substituting $q = 1$ into Eq. 87, we have the following for standard normal data

$$\begin{aligned} \mathbb{E} \left(D_{ij}^{(1*)} \right) &= \frac{\mu_{D_{ij}^{(1)}}}{2\mu_{\max}^{(1)}(m)} \\ &= \frac{p}{\sqrt{\pi}\mu_{\max}^{(1)}(m)}, \end{aligned} \quad (96)$$

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 81.

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

$$\begin{aligned} \text{Var} \left(D_{ij}^{(1*)} \right) &= \frac{6\log(m)\sigma_{D_{ij}^{(1)}}^2}{\pi^2 + 24 \left[\mu_{\max}^{(1)} \right]^2 \log(m)} \\ &= \frac{12p(\pi - 2)\log(m)}{\pi \left(\pi^2 + 24 \left[\mu_{\max}^{(1)} \right]^2 \log(m) \right)}, \end{aligned} \quad (97)$$

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 81.

2.5.2 Standard uniform data

Substituting $q = 1$ into Eq. 95, we have the following for standard uniform data

$$\begin{aligned} \mathbb{E} \left(D_{ij}^{(1*)} \right) &= \frac{(m+1)\mu_{D_{ij}^{(1)}}}{m-1} \\ &= \frac{(m+1)p}{3(m-1)}. \end{aligned} \quad (98)$$

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

$$\begin{aligned} \text{Var} \left(D_{ij}^{(1*)} \right) &= \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(1)}}^2}{m^3 - m + 2} \\ &= \frac{(m+2)(m+1)^2p}{18(m^3 - m + 2)}. \end{aligned} \quad (99)$$

2.6 Normalized Euclidean ($q = 2$)

Analogous to the previous section, we demonstrate the usage of Eqs. 87 and 95 for the max-min normalized Euclidean ($q = 2$) metric for both standard normal and standard uniform data.

2.6.1 Standard normal data

Substituting $q = 2$ into Eq. 87, we have the following for standard normal data

$$\begin{aligned} \mathbb{E} \left(D_{ij}^{(2*)} \right) &= \frac{\mu_{D_{ij}^{(2)}}}{2\mu_{\max}^{(1)}(m)} \\ &= \frac{\sqrt{2p-1}}{2\mu_{\max}^{(1)}(m)}, \end{aligned} \quad (100)$$

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 81. 366

Similarly, the variance of $D_{ij}^{(2*)}$ is given by 367

$$\begin{aligned}\text{Var}\left(D_{ij}^{(2*)}\right) &= \frac{6\log(m)\sigma_{D_{ij}^{(2)}}^2}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m)\right]^2\log(m)} \\ &= \frac{6\log(m)}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m)\right]^2\log(m)},\end{aligned}\tag{101}$$

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 81. 368

2.6.2 Standard uniform data 369

Substituting $q = 2$ into Eq. 95, we have the following for standard uniform data 370

$$\begin{aligned}\mathbb{E}\left(D_{ij}^{(2*)}\right) &= \frac{(m+1)\mu_{D_{ij}^{(2)}}}{m-1} \\ &= \sqrt{\frac{p}{6} - \frac{7}{120}} \left(\frac{m+1}{m-1}\right).\end{aligned}\tag{102}$$

Similarly, the variance of $D_{ij}^{(2*)}$ is given by 371

$$\begin{aligned}\text{Var}\left(D_{ij}^{(2*)}\right) &= \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(2)}}^2}{m^3 - m + 2} \\ &= \frac{7(m+2)(m+1)^2}{120(m^3 - m + 2)}.\end{aligned}\tag{103}$$

2.7 GWAS distance distributions 372

Consider a GWAS data set, which has the following encoding based on minor allele frequency 373

$$X_{ia} = \begin{cases} 0 & \text{if there are no minor alleles at locus } a, \\ 1 & \text{if there is 1 minor allele at locus } a, \\ 2 & \text{if there are 2 minor alleles at locus } a. \end{cases}\tag{104}$$

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 a . Furthermore, the minor allele probabilities are assumed to be independent and identically distributed according to $\mathcal{U}(l, u)$, where l and u are the lower and upper bounds, respectively, of the sampling distribution's support. 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 by 375

$$d_{ij}^{\text{GM}}(a) = \begin{cases} 0 & \text{if } X_{ia} \neq X_{ja}, \\ 1 & \text{otherwise} \end{cases}\tag{105}$$

and 384

$$d_{ij}^{\text{AM}}(a) = \frac{1}{2}|X_{ia} - X_{ja}|.\tag{106}$$

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 [9]. This metric is given by the following

$$d_{ij}^{\text{TiTv}}(a) = \begin{cases} 0 & \text{if } X_{ia} = X_{ja} \text{ and Ti/Tv,} \\ 1/4 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Ti,} \\ 1/2 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Tv,} \\ 3/4 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Ti,} \\ 1 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Tv.} \end{cases} \quad (107)$$

With any of the three metrics given by Eqs. 105 - 107, 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. 105 - 107 are asymptotically normal. Thus, we will proceed by deriving the mean and variance for each distance distribution induced by these three GWAS metrics.

2.7.1 GM distance distribution

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

$$\begin{aligned} \mathbb{E}[d_{ij}^{\text{GM}}(a)] &= \sum_{k=0}^1 k \cdot \mathbb{P}[d_{ij}^{\text{GM}}(a) = k] \\ &= 0 \cdot \mathbb{P}[d_{ij}^{\text{GM}}(a) = 0] + 1 \cdot \mathbb{P}[d_{ij}^{\text{GM}}(a) = 1] \\ &= \mathbb{P}[d_{ij}^{\text{GM}}(a) = 1] \\ &= 2\mathbb{P}[X_{ia} = 0, X_{ja} = 1] + 2\mathbb{P}[X_{ia} = 1, X_{ja} = 2] + 2\mathbb{P}[X_{ia} = 0, X_{ja} = 2] \\ &= 4(1 - f_a)^3 f_a + 4(1 - f_a) f_a^3 + 2(1 - f_a)^2 f_a^2 \\ &= 2[2(1 - f_a)^3 f_a + 2(1 - f_a) f_a^3 + (1 - f_a)^2 f_a^2] \\ &= 2F(a), \end{aligned} \quad (108)$$

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

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

$$\begin{aligned} \mathbb{E}\left(\sum_{a \in \mathcal{A}} d_{ij}^{\text{GM}}(a)\right) &= \sum_{a \in \mathcal{A}} \mathbb{E}[d_{ij}^{\text{GM}}(a)] \\ &= 2 \sum_{a \in \mathcal{A}} F(a). \end{aligned} \quad (109)$$

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

402

$$\begin{aligned}
\mathbb{E}[(D_{ij})^2] &= \mathbb{E}\left[\left(\sum_{a \in \mathcal{A}} d_{ij}^{\text{GM}}(a)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{a \in \mathcal{A}} (d_{ij}^{\text{GM}}(a))^2\right] + 2\mathbb{E}\left[\sum_{r \in \mathcal{A}} \sum_{s \leq r-1} d_{ij}^{\text{GM}}(r) \cdot d_{ij}^{\text{GM}}(s)\right] \\
&= \sum_{a \in \mathcal{A}} \left(\sum_{k=0}^1 k^2 \cdot \mathbb{P}[d_{ij}^{\text{GM}}(a) = k]\right) \\
&\quad + 2 \sum_{a \in \mathcal{A}} \sum_{s \leq r-1} \left(\sum_{k=0}^1 k \cdot \mathbb{P}[d_{ij}^{\text{GM}}(r) = k]\right) \cdot \left(\sum_{k=0}^1 k \cdot \mathbb{P}[d_{ij}^{\text{GM}}(s) = k]\right) \\
&= 2 \sum_{a \in \mathcal{A}} F(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} F(\lambda),
\end{aligned} \tag{110}$$

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

403

Using the moments given by Eqs. 109 and 110, the variance is computed as follows

404

$$\begin{aligned}
\text{Var}(D_{ij}) &= \mathbb{E}[(D_{ij})^2] - [\mathbb{E}(D_{ij})]^2 \\
&= 2 \sum_{a \in \mathcal{A}} F(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} F(\lambda) - 4 \left(\sum_{a \in \mathcal{A}} F(a)\right)^2 \\
&= 2 \sum_{a \in \mathcal{A}} F(a) - 4 \sum_{a \in \mathcal{A}} F^2(a) \\
&= 2 \sum_{a \in \mathcal{A}} F(a)[1 - 2F(a)],
\end{aligned} \tag{111}$$

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

405

With the mean and variance estimates given by Eqs. 109 and 111, the asymptotic GM distance distribution is given by the following

406

407

$$D_{ij} \sim \mathcal{N}\left(2 \sum_{a \in \mathcal{A}} F(a), 2 \sum_{a \in \mathcal{A}} F(a)[1 - 2F(a)]\right), \tag{112}$$

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

408

2.7.2 AM distance distribution

409

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

410

$$\begin{aligned}
\mathbb{E} [d_{ij}^{\text{AM}}(a)] &= \sum_{k \in \mathcal{D}} k \cdot \mathbb{P} [d_{ij}^{\text{AM}}(a) = k] \\
&= 0 \cdot \mathbb{P} [d_{ij}^{\text{AM}}(a) = 0] + \frac{1}{2} \cdot \mathbb{P} \left[d_{ij}^{\text{AM}}(a) = \frac{1}{2} \right] + 1 \cdot \mathbb{P} [d_{ij}^{\text{AM}}(a) = 1] \\
&= \frac{1}{2} (2\mathbb{P} [X_{ia} = 0, X_{ja} = 1] + 2\mathbb{P} [X_{ia} = 1, X_{ja} = 2]) \\
&\quad + 2\mathbb{P} [X_{ia} = 0, X_{ja} = 2] \\
&= \mathbb{P} [X_{ia} = 0, X_{ja} = 1] + \mathbb{P} [X_{ia} = 1, X_{ja} = 2] + 2\mathbb{P} [X_{ia} = 0, X_{ja} = 2] \\
&= 2(1 - f_a)^3 f_a + 2(1 - f_a) f_a^3 + 2(1 - f_a)^2 f_a^2 \\
&= 2[(1 - f_a)^3 f_a + (1 - f_a) f_a^3 + (1 - f_a)^2 f_a^2] \\
&= 2F(a),
\end{aligned} \tag{113}$$

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

411

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

412

413

$$\begin{aligned}
\mathbb{E} \left(\sum_{a \in \mathcal{A}} d_{ij}^{\text{AM}}(a) \right) &= \sum_{a \in \mathcal{A}} \mathbb{E} [d_{ij}^{\text{AM}}(a)] \\
&= 2 \sum_{a \in \mathcal{A}} F(a).
\end{aligned} \tag{114}$$

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

414

$$\begin{aligned}
\mathbb{E} [(D_{ij})^2] &= \mathbb{E} \left[\left(\sum_{a \in \mathcal{A}} d_{ij}^{\text{AM}}(a) \right)^2 \right] \\
&= \mathbb{E} \left[\sum_{a \in \mathcal{A}} (d_{ij}^{\text{AM}}(a))^2 \right] + 2\mathbb{E} \left[\sum_{r \in \mathcal{A}} \sum_{s \leq r-1} d_{ij}^{\text{AM}}(r) \cdot d_{ij}^{\text{AM}}(s) \right] \\
&= \sum_{a \in \mathcal{A}} \left(\sum_{k \in \mathcal{D}} k^2 \cdot \mathbb{P} [d_{ij}^{\text{AM}}(a) = k] \right) \\
&\quad + 2 \sum_{a \in \mathcal{A}} \sum_{s \leq r-1} \left(\sum_{k \in \mathcal{D}} k \cdot \mathbb{P} [d_{ij}^{\text{AM}}(r) = k] \right) \cdot \left(\sum_{k \in \mathcal{D}} k \cdot \mathbb{P} [d_{ij}^{\text{AM}}(s) = k] \right) \\
&= \sum_{a \in \mathcal{A}} G(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} F(\lambda),
\end{aligned} \tag{115}$$

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

415

416

Using the moments given by Eqs. 114 and 115, the variance is computed as follows 417

$$\begin{aligned}
\text{Var}(D_{ij}) &= E[(D_{ij})^2] - [E(D_{ij})]^2 \\
&= \sum_{a \in \mathcal{A}} G(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r,s\}} F(\lambda) - 4 \left(\sum_{a \in \mathcal{A}} F(a) \right)^2 \\
&= \sum_{a \in \mathcal{A}} G(a) - 4 \sum_{a \in \mathcal{A}} F^2(a) \\
&= \sum_{a \in \mathcal{A}} [G(a) - 4F^2(a)],
\end{aligned} \tag{116}$$

where $G(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a) + 2(1 - f_a)^2 f_a^2$ and $F(\lambda) = (1 - f_\lambda)^3 f_\lambda + f_\lambda^3 (1 - f_\lambda) + (1 - f_\lambda)^2 f_\lambda^2$. 418 419

With the mean and variance estimates given by Eqs. 114 and 116, the asymptotic AM distance distribution is given by the following 420 421

$$D_{ij} \sim \mathcal{N} \left(2 \sum_{a \in \mathcal{A}} F(a), \sum_{a \in \mathcal{A}} [G(a) - 4F^2(a)] \right), \tag{117}$$

where $G(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a) + 2(1 - f_a)^2 f_a^2$ and $F(\lambda) = (1 - f_\lambda)^3 f_\lambda + f_\lambda^3 (1 - f_\lambda) + (1 - f_\lambda)^2 f_\lambda^2$. 422 423

2.7.3 TiTv distance distribution 424

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 condition to compute the TiTv metric between instances i and 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). A diagram showing possible transitions and transversions that may occur is given by Fig. 3. Purines (A and G) and pyrimidines (C and T) are shown at the top and bottom, respectively. Transitions occur in the cases of PuPu and PyPy, while transversion occur only with PuPy encoding. 425 426 427 428 429 430 431 432 433

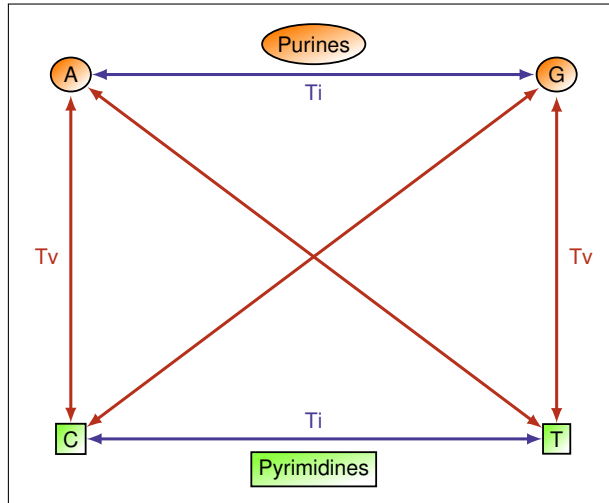


Fig 3. Purines (A and G) and pyrimidines (C and T) are shown. Transitions occur when a mutation involves purine-to-purine or pyrimidine-to-pyrimidine insertion. Transversions occur when a purine-to-pyrimidine or pyrimidine-to-purine insertion happens, which is a more extreme case. There are visibly more possibilities for transversions to occur than there are transitions, but there are about twice as many transitions in real data. 434

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. That is, the probability of a transition $P(\text{Ti})$ is approximately twice the probability of transversion $P(\text{Tv})$. It is likely that any particular data set will not satisfy this criterion exactly. In this general case, we have $P(\text{Ti})$ being equal to some multiple η times $P(\text{Tv})$. In order to enforce this general constraint in simulated data, we define the following set of equalities

$$\gamma_0 + \gamma_1 + \gamma_2 = 1, \quad (118)$$

$$P(\text{Ti}) - \eta P(\text{Tv}) = 0. \quad (119)$$

Using this PuPu, PuPy, and PyPy encoding, the probability of a transversion occurring at any fixed locus a is given by the following 435
436

$$P(\text{Tv}) = \gamma_1. \quad (120)$$

Using the constraints given by Eqs. 118 and 119, the probability of a transition occurring at locus a is computed as follows 437
438

$$P(\text{Ti}) = \gamma_0 + \gamma_2. \quad (121)$$

Also based on the constraints given by Eqs. 118 and 119, it is clear that we have 439
 $P(\text{Tv}) = \frac{1}{\eta+1}$ and $P(\text{Ti}) = \frac{\eta}{\eta+1}$. Without loss of generality, we then sample 440

$$\gamma_0 \sim \mathcal{U}\left(\varepsilon, \frac{\eta}{\eta+1} - \varepsilon\right), \quad (122)$$

where ε is some small positive real number. 441

Then it immediately follows that we have 442

$$\gamma_2 = \frac{\eta}{\eta+1} - \gamma_0. \quad (123)$$

However, we can derive the mean and variance of the distance distribution induced by the TiTv metric without specifying any relationship between γ_0 , γ_1 , and γ_2 . We proceed by computing $P[d_{ij}^{\text{TiTv}}(a) = k]$ for each $k \in \mathcal{D} = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. Let y represent a random sample of size p from $\{0, 1, 2\}$, where 443
444
445
446

$$y_a = \begin{cases} 0 & \text{if locus } a \text{ is PuPu,} \\ 1 & \text{if locus } a \text{ is PuPy,} \\ 2 & \text{if locus } a \text{ is PyPy.} \end{cases} \quad (124)$$

We derive $P[d_{ij}^{\text{TiTv}}(a) = 0]$ as follows 447

$$\begin{aligned} P[d_{ij}^{\text{TiTv}}(a) = 0] &= P[y_a = 0, X_{ia} = X_{ja}] \\ &\quad + P[y_a = 1, X_{ia} = X_{ja}] \\ &\quad + P[y_a = 2, X_{ia} = X_{ja}] \\ &= \gamma_0 [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &\quad + \gamma_1 [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &\quad + \gamma_2 [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &= (\gamma_0 + \gamma_1 + \gamma_2) [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &= (1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2. \end{aligned} \quad (125)$$

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

448

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

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

449

$$\begin{aligned}
P\left[d_{ij}^{\text{TiTv}}(a) = \frac{1}{2}\right] &= 2P[y_a = 1, X_{ia} = 0, X_{ja} = 1] \\
&\quad + 2P[y_a = 1, X_{ia} = 1, X_{ja} = 2] \\
&= 4\gamma_1(1-f_a)^3 f_a + 4\gamma_1 f_a^3(1-f_a) \\
&= 4\gamma_1 [(1-f_a)^3 f_a + f_a^3(1-f_a)].
\end{aligned} \tag{127}$$

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

450

$$\begin{aligned}
P\left[d_{ij}^{\text{TiTv}}(a) = \frac{3}{4}\right] &= 2P[y_a = 0, X_{ia} = 0, X_{ja} = 2] \\
&\quad + 2P[y_a = 2, X_{ia} = 0, X_{ja} = 2] \\
&= 2\gamma_0(1-f_a)^2 f_a^2 + 2\gamma_2(1-f_a)^2 f_a^2 \\
&= 2(\gamma_0 + \gamma_2)(1-f_a)^2 f_a^2.
\end{aligned} \tag{128}$$

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

451

$$\begin{aligned}
P[d_{ij}^{\text{TiTv}}(a) = 1] &= 2P[y_a = 1, X_{ia} = 0, X_{ja} = 2] \\
&= 2\gamma_1(1-f_a)^2 f_a^2.
\end{aligned} \tag{129}$$

Using Eqs. 125 - 129, we compute the expected TiTv distance between instances i and j as follows

452

453

$$\begin{aligned}
E(D_{ij}) &= \sum_{a \in \mathcal{A}} \left(\sum_{k \in \mathcal{D}} k \cdot P[d_{ij}^{\text{TiTv}}(a) = k] \right) \\
&= (\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}} [(1-f_a)^3 f_a + f_a^3(1-f_a)] \\
&\quad + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} (1-f_a)^2 f_a^2 \\
&= (\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G(a),
\end{aligned} \tag{130}$$

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

454

$$\begin{aligned}
\mathbb{E}[(D_{ij})^2] &= \mathbb{E}\left[\left(\sum_{a \in \mathcal{A}} d_{ij}^{\text{TiTv}}(a)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{a \in \mathcal{A}} (d_{ij}^{\text{TiTv}}(a))^2\right] + 2\mathbb{E}\left[\sum_{r \in \mathcal{A}} \sum_{s \leq r-1} d_{ij}^{\text{TiTv}}(r) \cdot d_{ij}^{\text{TiTv}}(s)\right] \\
&= \sum_{a \in \mathcal{A}} \left(\sum_{k \in \mathcal{D}} k^2 \cdot \mathbb{P}[d_{ij}^{\text{TiTv}}(a) = k]\right) \\
&\quad + 2 \sum_{a \in \mathcal{A}} \sum_{s \leq r-1} \left(\sum_{k \in \mathcal{D}} k \cdot \mathbb{P}[d_{ij}^{\text{TiTv}}(r) = k]\right) \cdot \left(\sum_{k \in \mathcal{D}} k \cdot \mathbb{P}[d_{ij}^{\text{TiTv}}(s) = k]\right) \\
&= \left[\frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1\right] \sum_{a \in \mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a) \\
&\quad + 2 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} \left([\gamma_0 + \gamma_2 + 2\gamma_1]F(\lambda) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right]G(\lambda)\right), \tag{131}
\end{aligned}$$

where $F(\lambda) = (1 - f_\lambda)^3 f_\lambda + f_\lambda^3 (1 - f_\lambda)$ and $G(\lambda) = (1 - f_\lambda)^2 f_\lambda^2$. 456

Using the moments given by Eqs. 130 and 131, the variance is computed as follows 457

$$\begin{aligned}
\text{Var}(D_{ij}) &= \mathbb{E}[(D_{ij})^2] - [\mathbb{E}(D_{ij})]^2 \\
&= \left[\frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1\right] \sum_{a \in \mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a) \\
&\quad + 2 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} \left([\gamma_0 + \gamma_2 + 2\gamma_1]F(\lambda) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right]G(\lambda)\right) \\
&\quad - \left([\gamma_0 + \gamma_2 + 2\gamma_1] \sum_{a \in \mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a)\right)^2 \\
&= \left[\frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1\right] \sum_{a \in \mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a) \\
&\quad - \sum_{a \in \mathcal{A}} \left([\gamma_0 + \gamma_2 + 2\gamma_1]F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right]G(a)\right)^2, \tag{132}
\end{aligned}$$

where $F(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a)$ and $G(a) = (1 - f_a)^2 f_a^2$. 458

With the mean and variance estimates given by Eqs. 130 and 132, the asymptotic TiTv distance distribution is given by the following 459 460

$$\begin{aligned}
D_{ij} &\sim \mathcal{N}\left((\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a), \right. \\
&\quad \left[\frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1\right] \sum_{a \in \mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a) \\
&\quad \left. - \sum_{a \in \mathcal{A}} \left([\gamma_0 + \gamma_2 + 2\gamma_1]F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right]G(a)\right)^2\right), \tag{133}
\end{aligned}$$

where $F(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a)$ and $G(a) = (1 - f_a)^2 f_a^2$. 461

The relationship between the average success probability \bar{f}_a and the predicted TiTv pairwise distance given by Eq. 130 is shown in Fig. 4. Given upper and lower bounds l and u , respectively, of the success probability sampling interval, the average success probability (or average MAF) is computed as follows

$$\bar{f}_a = \frac{1}{2}(l + u). \quad (134)$$

The maximum distance occurs at $\bar{f}_a = 0.5$, which is the inflection point about which the minor allele changes at locus a . If few minor alleles are present ($\bar{f}_a \rightarrow 0$), the predicted TiTv distance approaches 0. The same is true after the minor allele switches ($\bar{f}_a \rightarrow 1$).

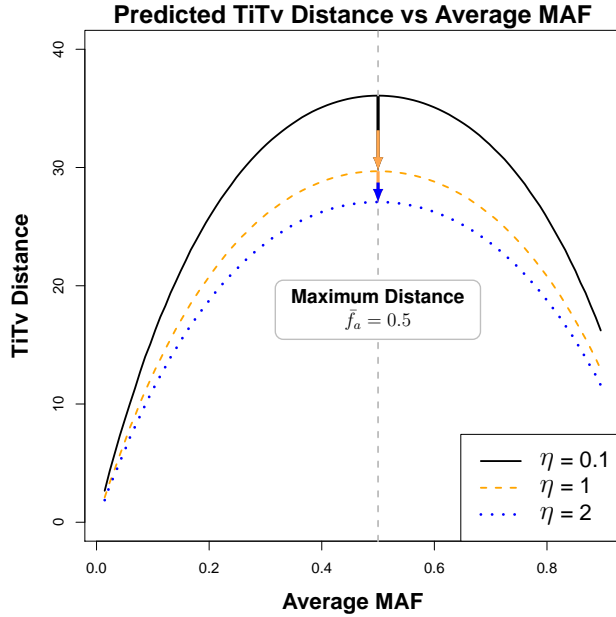


Fig 4. Predicted average TiTv distance as a function of average minor allele frequency \bar{f}_a (see Eq. 134). Success probabilities f_a were drawn from a sliding window interval from 0.01 to 0.9 in increments of about 0.009. With $\eta = 0.1$, where η is the Ti/Tv ratio given by Eq. 118, Tv is ten times more likely than Ti so the distance is large. Increasing to $\eta = 1$, Tv and Ti are equally likely so the distance is moderate. In line with real data for $\eta = 2$, Tv is half as likely as Ti so the distance is relatively small.

2.8 Time series correlation-based 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

$$d_{ij}^{\text{ROI}}(a) = \sum_{k \neq a} |A_{ka}^{(i)} - A_{ka}^{(j)}|. \quad (135)$$

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 (see Fig. 5).

$$\begin{array}{c}
\text{ROI}_1 \left\{ \begin{array}{c} \hat{A}_{12}^{(1)} \quad \hat{A}_{12}^{(2)} \quad \hat{A}_{12}^{(3)} \quad \hat{A}_{12}^{(4)} \quad \dots \quad \hat{A}_{12}^{(m)} \\ \hat{A}_{13}^{(1)} \quad \hat{A}_{13}^{(2)} \quad \hat{A}_{13}^{(3)} \quad \hat{A}_{13}^{(4)} \quad \dots \quad \hat{A}_{13}^{(m)} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \dots \quad \vdots \\ \hat{A}_{1p}^{(1)} \quad \hat{A}_{1p}^{(2)} \quad \hat{A}_{1p}^{(3)} \quad \hat{A}_{1p}^{(4)} \quad \dots \quad \hat{A}_{1p}^{(m)} \end{array} \right. \\
\text{ROI}_2 \left\{ \begin{array}{c} \hat{A}_{21}^{(1)} \quad \hat{A}_{21}^{(2)} \quad \hat{A}_{21}^{(3)} \quad \hat{A}_{21}^{(4)} \quad \dots \quad \hat{A}_{21}^{(m)} \\ \hat{A}_{23}^{(1)} \quad \hat{A}_{23}^{(2)} \quad \hat{A}_{23}^{(3)} \quad \hat{A}_{23}^{(4)} \quad \dots \quad \hat{A}_{23}^{(m)} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \dots \quad \vdots \\ \hat{A}_{2p}^{(1)} \quad \hat{A}_{2p}^{(2)} \quad \hat{A}_{2p}^{(3)} \quad \hat{A}_{2p}^{(4)} \quad \dots \quad \hat{A}_{2p}^{(m)} \end{array} \right. \\
\vdots \\
\text{ROI}_p \left\{ \begin{array}{c} \hat{A}_{p1}^{(1)} \quad \hat{A}_{p1}^{(2)} \quad \hat{A}_{p1}^{(3)} \quad \hat{A}_{p1}^{(4)} \quad \dots \quad \hat{A}_{p1}^{(m)} \\ \hat{A}_{p2}^{(1)} \quad \hat{A}_{p2}^{(2)} \quad \hat{A}_{p2}^{(3)} \quad \hat{A}_{p2}^{(4)} \quad \dots \quad \hat{A}_{p2}^{(m)} \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \dots \quad \vdots \\ \hat{A}_{p,p-1}^{(1)} \quad \hat{A}_{p,p-1}^{(2)} \quad \hat{A}_{p,p-1}^{(3)} \quad \hat{A}_{p,p-1}^{(4)} \quad \dots \quad \hat{A}_{p,p-1}^{(m)} \end{array} \right.
\end{array} = X$$

Fig 5. Resting-state fMRI transformed subject correlation matrices. Each column corresponds to an instance (or subject) I_j and each column corresponds to an ROI (or feature). The notation $\hat{A}_{ka}^{(j)}$ represents the r-to-z transformed correlation between ROIs a and $k \neq a$ for instance j .

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. 43 and 44, 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

$$\begin{aligned}
E(D_{ij}) &= E \left(\sum_{a \in \mathcal{A}} \sum_{k \neq a} |\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}| \right) \\
&= \sum_{a \in \mathcal{A}} \sum_{k \neq a} E \left(|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}| \right) \\
&= \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{2}{\sqrt{\pi}} \\
&= \frac{2p(p-1)}{\sqrt{\pi}}.
\end{aligned} \tag{136}$$

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

489

$$\begin{aligned}
\text{Var}(D_{ij}) &= \text{Var} \left(\sum_{a \in \mathcal{A}} \sum_{k \neq a} |\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}| \right) \\
&= \sum_{a=1}^{p-1} \text{Var} \left(\sum_{k=a+1}^p 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}| \right) \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left(\sum_{k=a+1}^p 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^p 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}| \right) \\
&= \sum_{a=1}^{p-1} \sum_{k=a+1}^p \text{Var} \left(2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}| \right) \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left(\sum_{k=a+1}^p 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^p 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}| \right) \quad (137) \\
&= \sum_{a=1}^{p-1} \sum_{k=a+1}^{p-1} \frac{4(\pi - 2)}{\pi} \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left(\sum_{k=a+1}^p 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^p 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}| \right) \\
&= \frac{2p(\pi - 2)(p - 1)}{\pi} \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left(\sum_{k=a+1}^p 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^p 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}| \right).
\end{aligned}$$

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. 137. Through simulation, it can be seen that the difference between the sample variance $S_{D_{ij}}^2$ and $\frac{2p(\pi-2)(p-1)}{\pi}$ has a quadratic relationship with p . More explicitly, we have the following relationship

490
491
492
493

$$S_{D_{ij}}^2 - \frac{2p(\pi - 2)(p - 1)}{\pi} = \beta_1 p^2 + \beta_0 p. \quad (138)$$

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

494
495
496
497

$$2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left(\sum_{k=a+1}^p 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^p 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}| \right) = \frac{p(\pi - 2)(p - 1)}{4\pi}. \quad (139)$$

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

498

$$\text{Var}(D_{ij}) = \frac{9p(\pi - 2)(p - 1)}{4\pi}. \quad (140)$$

With the mean and variance estimates given by Eqs. 136 and 140, we have the following asymptotic distribution for rs-fMRI distances

499
500

$$D_{ij}^{(1)} \sim \mathcal{N} \left(\frac{2p(p - 1)}{\sqrt{\pi}}, \frac{9p(\pi - 2)(p - 1)}{4\pi} \right). \quad (141)$$

Consider the max-min normalized rs-fMRI distance given by the following equation 501

$$D_{ij}^{1*} = \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{|A_{ak}^{(i)} - A_{ak}^{(j)}|}{\max(a) - \min(a)}. \quad (142)$$

Assuming that the data X has been r-to-z transformed and standardized, we can 502
easily compute the expected attribute range and variance of the attribute range. 503
The expected maximum of a given attribute in data matrix X is estimated by the following 504

$$\mathbb{E}(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m, p) = 2 \left[\frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m(p-1)}\right)} - \Phi^{-1}\left(\frac{1}{m(p-1)}\right) \right]. \quad (143)$$

The variance can be esimated with the following 505

$$\text{Var}(X_a^{\max} - X_a^{\min}) = \frac{\pi^2}{6 \log[m(p-1)]}. \quad (144)$$

Let $\mu_{D_{ij}}$ and $\sigma_{D_{ij}}^2$ denote the mean and variance of the rs-fMRI distance distribution 506
given by Eqs. 136 and 140. Using the formulas for the mean and variance of the max-min 507
normalized distance distribution given in Eq. 87, we have the following asymptotic 508
distribution for the max-min normalized rs-fMRI distances 509

$$D_{ij}^{1*} \sim \mathcal{N} \left(\frac{\mu_{D_{ij}}^{(1)}}{2\mu_{\max}^{(1)}(m, p)}, \frac{6\sigma_{D_{ij}}^{(1)2} \log[m(p-1)]}{\pi^2 + 24 [\mu_{\max}^{(1)}(m, p)]^2 \log[m(p-1)]} \right). \quad (145)$$

2.9 Normalized Manhattan ($q = 1$) for rs-fMRI 510

Substituting the non-normalized mean given by Eq. 136 into Eq. 145 for the mean of 511
the max-min normalized rs-fMRI metric, we have the following 512

$$\begin{aligned} \mathbb{E}(D_{ij}^{(1*)}) &= \frac{\mu_{D_{ij}}^{(1)}}{2\mu_{\max}^{(1)}(m, p)} \\ &= \frac{p(p-1)}{\sqrt{\pi} \mu_{\max}^{(1)}(m, p)}, \end{aligned} \quad (146)$$

where $\mu_{\max}^{(1)}(m, p)$ is given in Eq. 143. 513

Similarly, the variance of $D_{ij}^{(1*)}$ is given by 514

$$\begin{aligned} \text{Var}(D_{ij}^{(1*)}) &= \frac{6\sigma_{D_{ij}}^{(1)2} \log[m(p-1)]}{\pi^2 + 24 [\mu_{\max}^{(1)}(m, p)]^2 \log[m(p-1)]} \\ &= \frac{27(\pi - 2) \log[m(p-1)](p-1)p}{2\pi \left(\pi^2 + 24 [\mu_{\max}^{(1)}(m, p)]^2 \log[m(p-1)] \right)}, \end{aligned} \quad (147)$$

where $\mu_{\max}^{(1)}(m, p)$ is given in Eq. 143. 515

3 Effects of correlation on distances

All of the derivations presented in previous sections are for the cases where there is no correlation between instances or features. We assumed that any pair (X_{ia}, X_{ja}) of data points for instances i and j and fixed feature a were independent and identically distributed. This was done in order to determine asymptotic estimates in null data. That is, data with no main effects, interaction effects, or pairwise correlations between features. Within this highly simplified context, our asymptotic formulas for distributional moments are reliable. However, correlations do exist between features and instances in real data. There are a multitude of different statistical effects that impact distance distributional properties. Ultimately, divergence from normality is caused primarily by large magnitude pairwise correlation between features. Pairwise feature correlation can be the result of main effects, where features have different within-group means. On the other hand, there could be an underlying interaction network in which there are strong associations between features. If features are differentially correlated between phenotype groups, then interactions exist that change affect the distance distribution. In the following few sections, we consider particular cases of the L_q metric for continuous and discrete data under the effects of pairwise feature correlation.

3.1 Continuous data

Consider $X^{(m \times p)}$ where $X_{ia} \sim \mathcal{N}(0, 1)$ for all $i = 1, 2, \dots, m$ and $a = 1, 2, \dots, p$. Without loss of generality, we let $m = p = 100$ and consider only the L_2 (Euclidean) metric. An illustration of the effects of correlation on distances with the given assumptions is shown in Fig. 6. Each density curve shown in (blue) is for a simulated distance matrix from data with some degree of pairwise correlation between features. Divergence from normality in distances is directly related to the average absolute pairwise correlation that exists in the simulated data. This measure is given by

$$\bar{r}_{\text{abs}} = \frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j>i} r_{ij} \quad (148)$$

where r_{ij} is the correlation between features i and j across all instances m . The distance density curve (orange) is representative of distances generated from random standard normal data with no added correlation. The mean and variance of this distribution are given by Eqs. 49 and 48, respectively, by substituting $p = 100$ for the mean. From left-to-right and top-to-bottom, there is an increase in \bar{r}_{abs} . This very quickly introduces positive skewness and increased variability. The predicted and sample means, however, are approximately the same in each case due to linearity of the expectation operator. Because of the dependencies between features, the predicted variance of 1 obviously no longer holds.

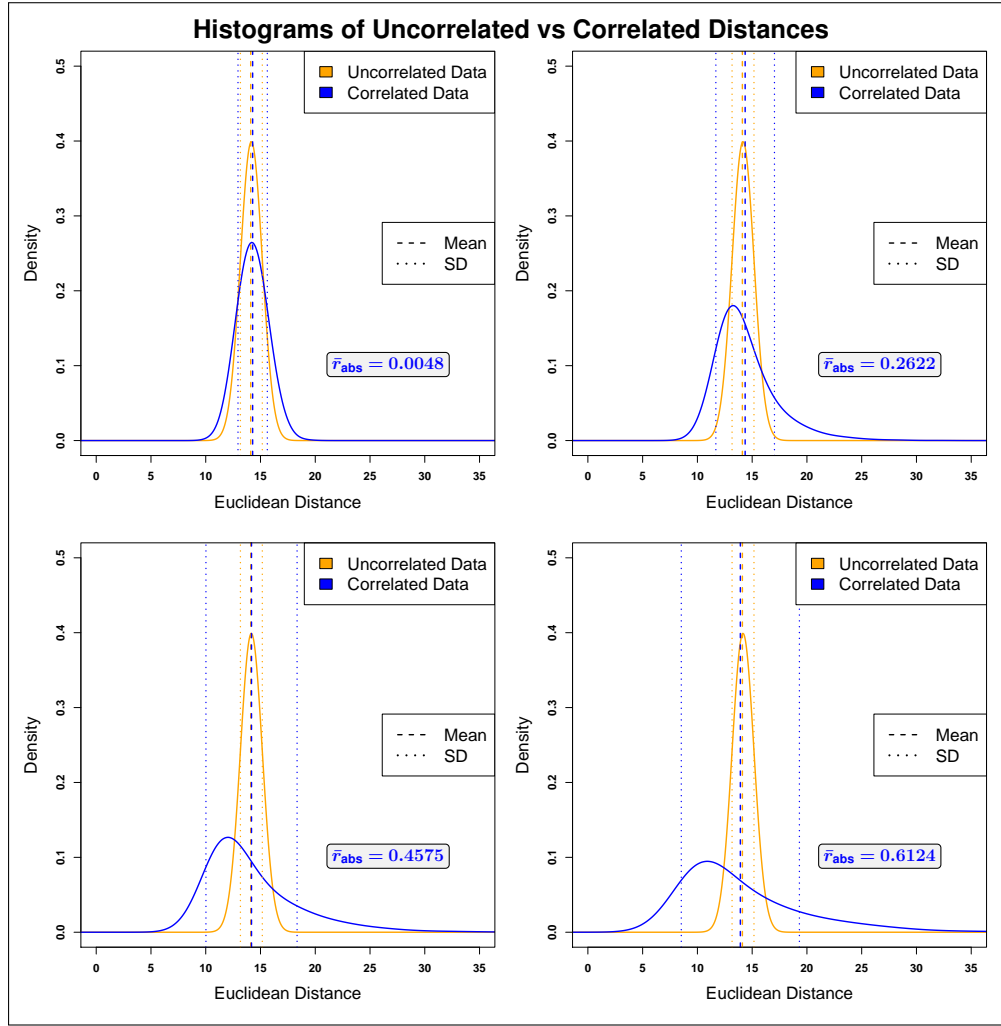


Fig 6. Density curves of distance distributions generated from correlated and null data from standard normal distribution. Each plot has the theoretical density curve for the null Euclidean distance distribution (orange) and a density curve for Euclidean distances from correlated data (blue) with some average magnitude pairwise correlation (\bar{r}_{abs}) between features. (**Upper left**) With little correlation, distances begin to diverge from the predicted normal distribution. (**Upper Right**) Low to moderate correlation causes a large positive skew and increased variability in distances. (**Lower Left**) Moderate to high correlation increases skewness and variability in distances. (**Lower Right**) Extreme correlation produces maximal skewness and variability in distances. In each case, the null and correlated distance means are approximately the same.

3.2 GWAS data

In analogy to the previous section, we explore the effects of pairwise feature correlation in the context of GWAS data. Without loss of generality, we let $m = p = 100$ and consider the TiTv metric, which is given by combining Eqs. 107 and 1 with $q = 1$.

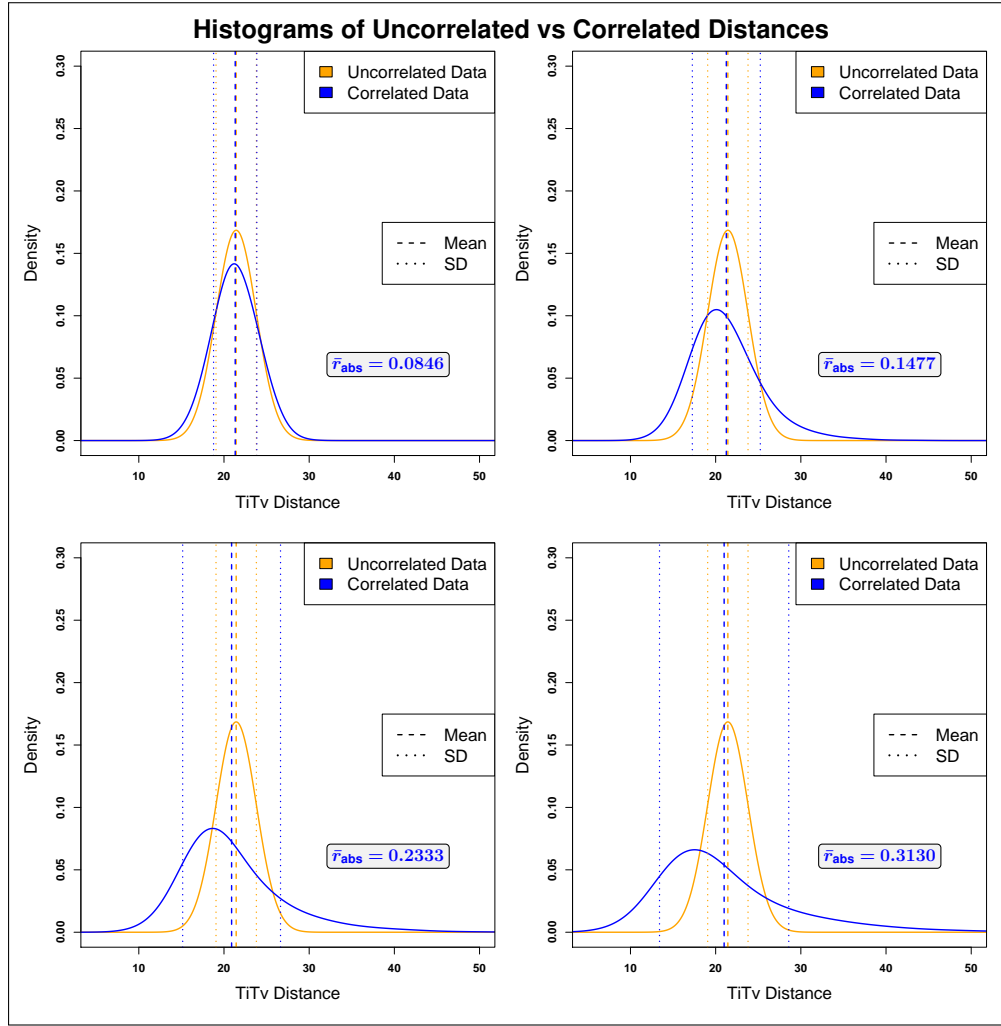


Fig 7. Density curves of distance distributions generated from correlated and null data from standard normal distribution. Each plot has the theoretical density curve for the null Euclidean distance distribution (orange) and a density curve for TiTv distances from correlated data (blue) with some average magnitude pairwise correlation (\bar{r}_{abs}) between features. (**Upper left**) With little correlation, distances begin to diverge from the predicted normal distribution. (**Upper Right**) Low to moderate correlation causes a large positive skew and increased variability in distances. (**Lower Left**) Moderate to high correlation increases skewness and variability in distances. (**Lower Right**) Extreme correlation produces maximal skewness and variability in distances. In each case, the null and correlated distance means are approximately the same.

3.3 Correlation-based data

For our correlation data-based metric given by Eqs. 135 and 1 with $q = 1$, we consider additional effects of correlation between features. Without loss of generality, we let $m = 100$ and $p = 30$.

554

555

556

557

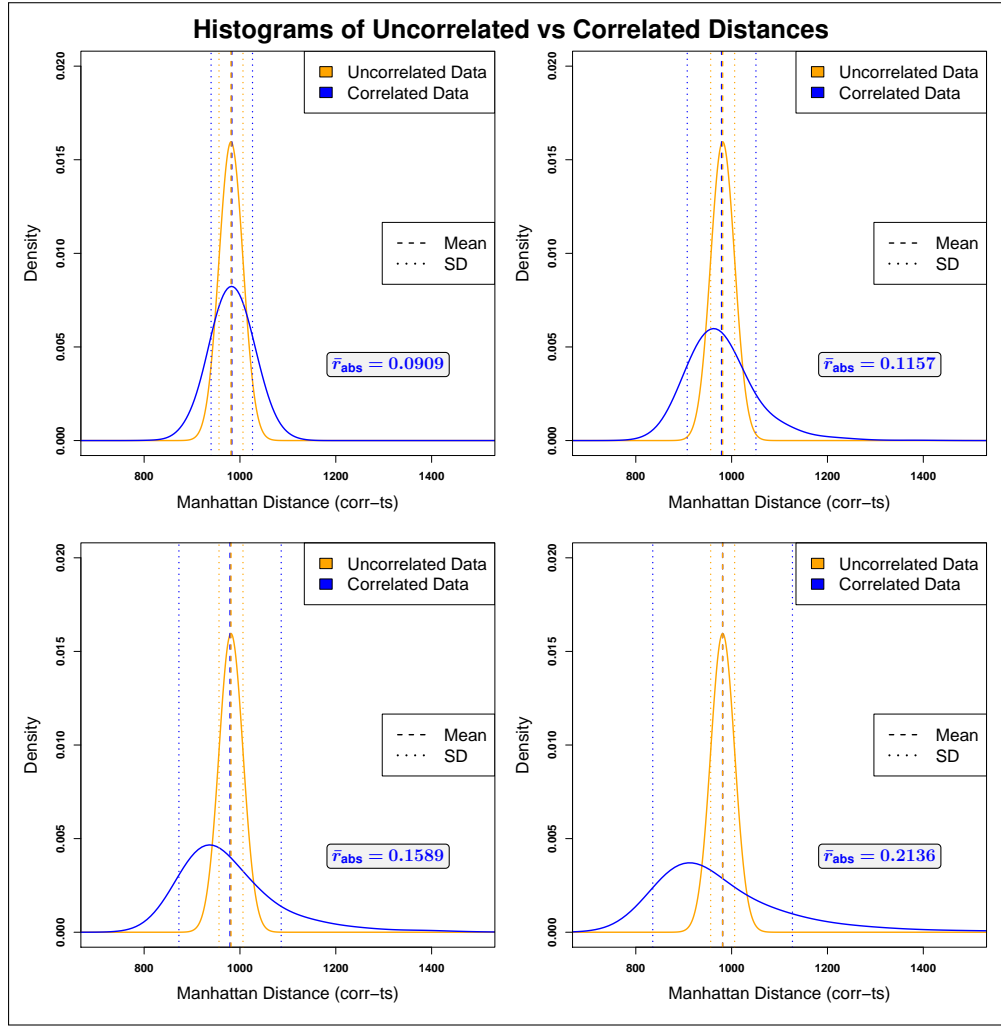


Fig 8. Density curves of distance distributions generated from correlated and null data from standard normal distribution. Each plot has the theoretical density curve for the null Euclidean distance distribution (orange) and a density curve for Manhattan distances from correlated corr-ts data (blue) with some average magnitude pairwise correlation (\bar{r}_{abs}) between features. (**Upper left**) With little correlation, distances begin to diverge from the predicted normal distribution. (**Upper Right**) Low to moderate correlation causes a large positive skew and increased variability in distances. (**Lower Left**) Moderate to high correlation increases skewness and variability in distances. (**Lower Right**) Extreme correlation produces maximal skewness and variability in distances. In each case, the null and correlated distance means are approximately the same.

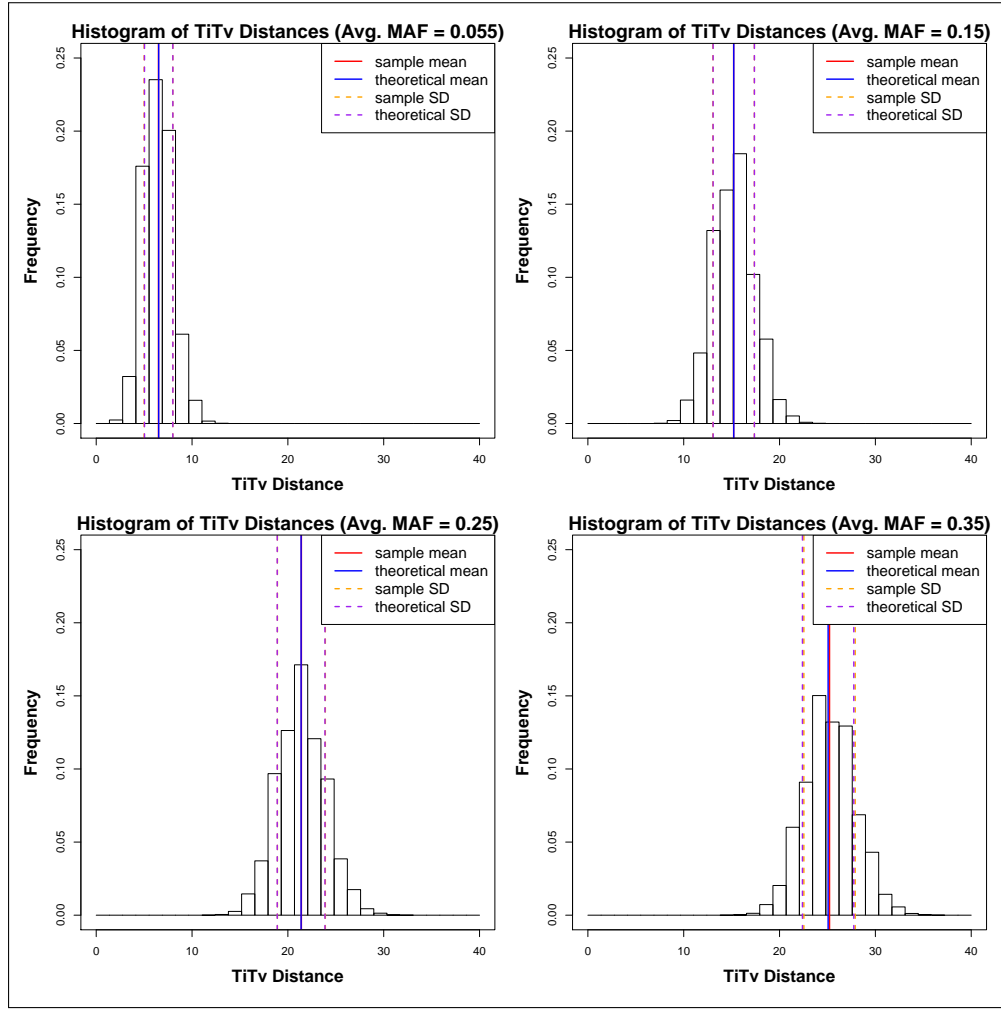


Fig 9. Histograms of simulated TiTv distance distributions for different average MAFs. The Ti/Tv ratio was fixed to be 2 in all simulations. Average MAF is computed as the expected value of the uniform distribution from which minor allele success probabilities (f_a) are drawn. The upper bounds for each success probability uniform distribution are $\{0.1, 0.2, 0.3, 0.4\}$, which are the maximum possible MAF for a given locus a . The corresponding lower bounds were $\{0.01, 0.1, 0.2, 0.3\}$. Sample and predicted means, as well as standard deviations, are overlaid on each histogram. Each distance distribution comes from a simulated data set with $m = 100$ instances and $p = 100$ features.

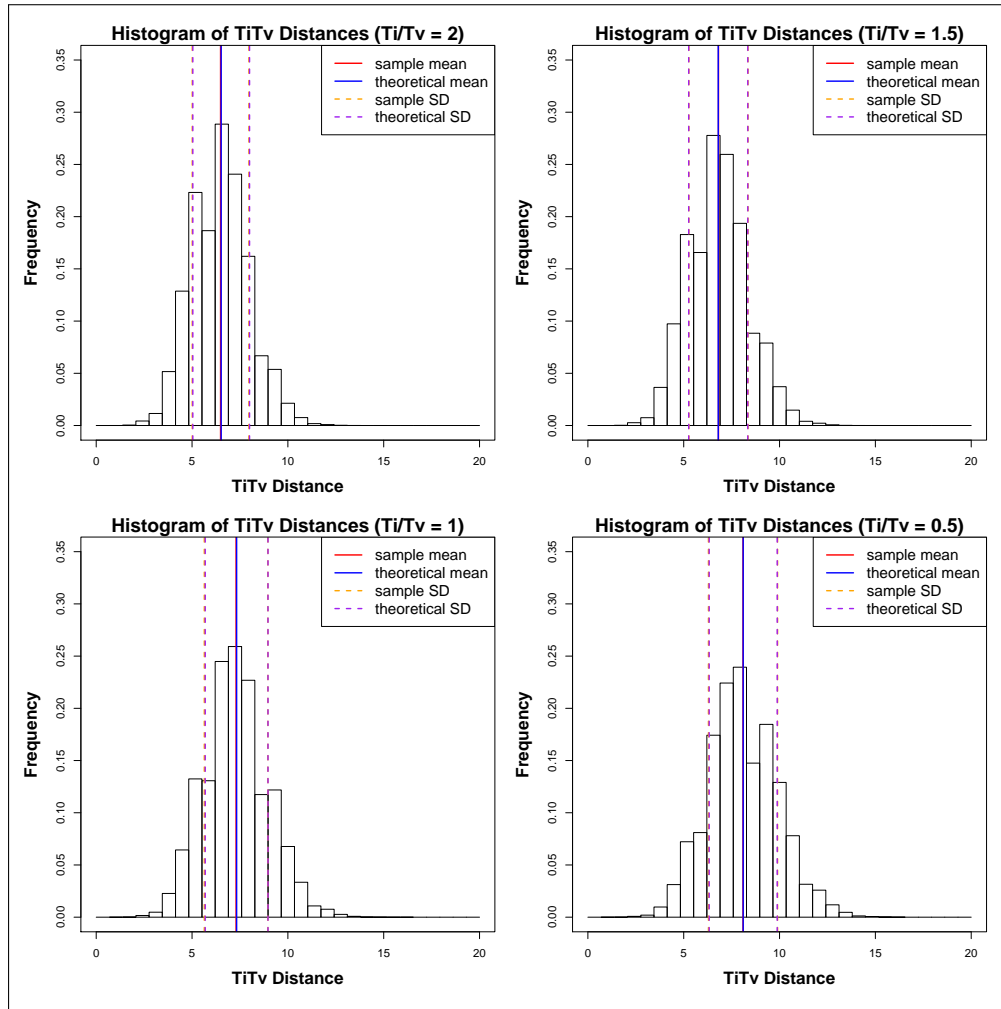


Fig 10. Histograms of simulated TiTv distance distributions for different Ti/Tv ratios. Average MAF was fixed to be 0.055. The Ti/Tv ratio was taken to be 2, 1.5, 1, and 0.5. The average distance increases as the Ti/Tv ratio decreases, which is intuitive because the TiTv distance is greater for transversions than transitions. Sample and predicted means, as well as standard deviations, are overlaid on each histogram. Each distance distribution comes from a simulated data set with $m = 100$ instances and $p = 100$ features.

Table 1. Summary of distance distribution derivations for standard normal and standard uniform data. Asymptotic estimates are given for both standard and max-min normalized q-metrics. These estimates are relevant for all $q \in \mathbb{N}$ and $p \geq 100$.

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

Table 2. Asymptotic estimates for means and variances for the standard L_1 and L_2 distance distributions. Estimates for both standard normal and standard uniform data are given.

q -Metric	Data	Stat	Formula (Eq. #)
standard (L_1)	$\mathcal{N}(0, 1)$	mean	$\frac{2p}{\sqrt{\pi}} \quad (38)$
		variance	$\frac{2(\pi-2)p}{\pi} \quad (38)$
	$\mathcal{U}(0, 1)$	mean	$\frac{p}{3} \quad (48)$
		variance	$\frac{p}{18} \quad (48)$
standard (L_2)	$\mathcal{N}(0, 1)$	mean	$\sqrt{2p-1} \quad (38)$
		variance	$1 \quad (38)$
	$\mathcal{U}(0, 1)$	mean	$\sqrt{\frac{p}{6} - \frac{7}{120}} \quad (48)$
		variance	$\frac{7}{120} \quad (48)$

Table 3. Asymptotic estimates for means and variances for the max-min normalized L_1 and L_2 distance distributions. Estimates for both standard normal and standard uniform data are given.

q -Metric	Data	Stat	Formula (Eq. #)
max-min normalized (L_1)	$\mathcal{N}(0, 1)$	mean	$\frac{p}{\sqrt{\pi}\mu_{\max}^{(1)}(m)} \quad (93)$ where $\mu_{\max}^{(1)}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
		variance	$\frac{12p(\pi-2)\log(m)}{\pi\left(\pi^2+24\left[\mu_{\max}^{(1)}(m)\right]^2\log(m)\right)} \quad (93)$ where $\mu_{\max}^{(1)}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
	$\mathcal{U}(0, 1)$	mean	$\frac{(m+1)p}{3(m-1)} \quad (101)$
		variance	$\frac{(m+2)(m+1)^2p}{18(m^3-m+2)} \quad (48)$
max-min normalized (L_2)	$\mathcal{N}(0, 1)$	mean	$\frac{\sqrt{2p-1}}{2\mu_{\max}^{(1)}(m)} \quad (93)$ where $\mu_{\max}^{(1)}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
		variance	$\frac{6\log(m)}{\pi^2+24\left[\mu_{\max}^{(1)}(m)\right]^2\log(m)} \quad (93)$ where $\mu_{\max}^{(1)}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
	$\mathcal{U}(0, 1)$	mean	$\sqrt{\frac{p}{6} - \frac{7}{120} \left(\frac{m+1}{m-1}\right)} \quad (101)$
		variance	$\frac{7(m+2)(m+1)^2}{120(m^3-m+2)} \quad (101)$

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

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

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

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

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. Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 2018.
3. Marko Robnik Šikonja and Igor Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 53:23 – 69, February 2003.
4. Archana Venkataraman, Marek Kubicki, Carl-Fredrik Westin, and Polina Golland. Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies. *Conf Comput Vis Pattern Recognit Workshops*, pages 63–70, 2010.
5. Etay Hay, Petra Ritter, Nancy J. Lobaugh, and Anthony R. McIntosh. Multiregional integration in the brain during resting-state fMRI activity. *PLOS Computational Biology*, March 2017.
6. Benedikt Sundermann, Mona Olde lütke Beverborg, and Bettina Pfleiderer. Toward literature-based feature selection for diagnostic classification: a meta-analysis of resting-state fMRI in depression. *Frontiers in Human Neuroscience*, September 2014.
7. Svyatoslav Vergun, Alok S. Deshpande, Timothy B. Meier, Jie Song, Dana L. Tudorascu, Veena A. Nair, Vikas Singh, Bharat B. Biswal, M. Elizabeth Meverand, Rasmus M. Birn, and Vivek Prabhakaran. Characterizing functional connectivity differences in aging adults using machine learning on resting state fMRI data. *Frontiers in Computational Neuroscience*, April 2013.

8. Trang T Le, Ryan J Urbanowicz, Jason H Moore, and Brett A McKinney. Statis- 582
tical inference relief (stir) feature selection. *Bioinformatics*, page bty788, 2018. 583
9. M. Arabnejad, B. A. Dawkins, W. S. Bush, B. C. White, A. R. Harkness, and 584
B. A. McKinney. Transition-transversion encoding and genetic relationship metflic 585
in ReliefF feature selection improves pathway enrichment in GWAS. *BioData* 586
Mining, 11(23), 2018. 587
10. Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. 588
Springer, New York, NY, 2004. 589
11. Irwin Miller and Marylees Miller. *John E. Freund's Mathematical Statistics with* 590
Applications. Pearson Prentice Hall, 7 edition, 2004. 591
12. Ryan J. Urbanowicz, Melissa Meeker, William LaCava, Randal S. Olson, and 592
Jason H. Moore. Relief-Based Feature Selection: Introduction and Review. 593
arXiv:1711.08421 [cs.DS], 2018. 594
13. E. J. Gumbel. The Distribution of the Range. *The Annals of Mathematical* 595
Statistics, 18(3):384–412, September 1947. 596
14. Sourav Chatterjee. *Superconcentration and Related Topics*. 1439-7382. Springer 597
International Publishing, 1 edition, 2014. 598
15. Harald Cramér. *Mathematical Methods of Statistics*, volume 1. Princeton University 599
Press, reprint, revised edition, 1999. 600