Theoretical properties of nearest-neighbor distance distributions and novel metrics for high dimensional bioinformatics data

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

Abstract

The performance of nearest-neighbor feature selection and prediction methods depends on the metric for computing neighborhoods and the distribution type of the underlying data. The effects of the distribution and metric, as well as the presence of correlation and interactions, are reflected in the expected moments of the distribution of pairwise distances. We derive general analytical expressions for the mean and variance of pairwise distances for L_q metrics for Gaussian and uniform data with p attributes and m instances. We use extreme value theory to derive results for metrics normalized by the max and min of attributes. These expressions are applicable to the analysis of continuous data such as gene expression. We derive similar analytical expressions for a new metric for genetic variants in GWAS data (categorical predictors) that accounts for minor allele frequency and transition/transversion ratio. We introduce a new metric for resting-state fMRI data that is applicable to correlation-based predictors from time series data. Derivations assume independent data, but empirically we also consider the effect of correlation. This study provides detailed derivations and expressions parameterized by q, p, m and other properties for a broad collection of bioinformatics data types.

Author summary

Introduction

Statistical models can deviate from expected behavior depending on whether certain properties of the underlying data are satisfied, such as having a Gaussian distribution. Nearest neighbor methods are further influenced by the choice of metric, such as Euclidean or Manhattan. For random normal data, for example, the variance of the pairwise distances of a Manhattan metric is proportional to the number of attributes (p) whereas the variance is constant for a Euclidean metric. Relief methods [1–3] and nearest-neighbor projected distance regression (NDPR) use nearest neighbors to compute attribute importance scores and often use adaptive neighborhoods that rely on the mean and variance of the distance distribution. Thus, knowledge of the expected values for a given metric and data distribution may improve the performance of these feature selection methods.

For continuous data, the metrics most commonly used in nearest neighbor methods are L_q with $q = 1$ (Manhattan) or $q = 2$ (Euclidean). 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 readily available in the literature. We provide detailed derivations

of generalized expressions parameterized by q, attributes p, and samples m. We build on this mathematical formalism to derive the asymptotic properties of a new metric for categorical data in genome-wide association studies (GWAS) data [4]. We also derive asymptotic properties of a new metric introduced in the current study for resting-state fMRI (rs-fMRI) correlation data.

We derive asymptotic formulas for the mean and variance for three recently introduced GWAS metrics [4]. These metrics were developed for Relief-based feature selection to account for binary variant differences (two levels), allelic differences (three levels), and transition/transversion differences (five levels). The mean and variance expressions we derive for these multi-level categorical data types are parameterized by the minor allele frequency and the transition/transversion ratio.

The analysis of rs-fMRI data is a growing area for machine learning and feature selection [5–8]. For a given subject, a time series correlation or similar matrix is computed between regions of interest (ROIs), consisting of subsets of voxels from the neuroimage data. The time series represent functional activity of the ROI while the subject is at rest and the ROI typically corresponds to a region with known function for emotion or cognition. Thus, the dataset consists of pairwise ROI correlations for each of the m subjects. Nearest-neighbor based feature selection was applied to rs-fMRI in the private evaporative cooling method [cite pec], where the predictors were pairwise correlations between ROIs. The use of pairwise correlation predictors is a common practice because of convenience and differential connectivity between brain regions may be of biological importance [9]. However, one may be interested in the importance of features at the ROI level. Thus, in the current study we introduce a new metric to be used in NPDR with resting state correlation matrices that provides feature importance for ROIs. This metric is applicable to general time series-correlation (ts-corr) based data, and we derive asymptotic estimates for the mean and variance of distance distributions induced by our new ts-corr based metric.

The ability of nearest neighbor feature selection to identify association effects, like main effects or interaction effects, depends on neighborhood parameters, such as neighborhood radius or number of neighbors k. As k increases, nearest neighbor distance based algorithms are more sensitive to detecting main effects [10]. On the other hand, their ability to detect interaction effects decreases with increasing k [10, 11]. Correlation and interaction effects impact distance distributions by introducing positive skewness and increased variance, which can lead to changes in neighborhood inclusion. In order to understand how these statistical effects impact distance distributions in continuous and discrete data types, we first derive distance asymptotics for independently and identically distributed data. Using these derivations as a baseline, we can then determine how statistical effects and correlation change distance distributional properties from the null case.

In Section 1, we introduce preliminary notation and apply the Central Limit Theorem (CLT) and the Delta Method to derive asymptotics for pairwise distances. In Section 2, we present general derivations for continuously distributed data sets with m instances and p features. We focus on the cases of standard normal and standard uniform data distributions, but we derive analytical expressions parameterized by p and q. In Section 2.4 we use Extreme Value Theory (EVT) to address max-min normalized versions of L_q metrics, which are often used in Relief-based algorithms. In Section 3, we extend the derivations to categorical data with a binomial distribution for GWAS data. In Section 4, we present the final set of asymptotic results for our newly introduced time series correlation-based distance metric, with a particular emphasis on rs-fMRI data. Lastly, in Section 6, we demonstrate the effect of correlation in the attribute space on distance distributional properties.

1 Limit distribution for L_q on null data

In the application of nearest-neighbor distance-based methods to continuous data, the distance between instances ($i, j \in \mathcal{I}, |\mathcal{I}| = m$) in the data set $X^{m\times p}$ of m instances and p attributes (or features) is calculated in the space of all attributes ($a \in \mathcal{A}$, $|\mathcal{A}| = p$) using a metric such as

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

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

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

where $\hat{X}$ represents the standardized data matrix X. We use a simplified $\mathrm{d}_{ij}(a)$ notation in place of the $\mathrm{diff}(a,(i,j))$ notation that is customary in Relief-based methods. In NPDR, we omit the division by $\max(a) - \min(a)$ used by Relief to constrain scores to the interval from -1 to 1, where $\max(a) = \max_{k\in\mathcal{I}}\{X_{ka}\}$ and $\min(a) = \min k \in \mathcal{I}\{X_{ka}\}$. The numeric $\mathrm{d}_{ij}^{\mathrm{num}}(a)$ projection is simply the absolute difference between row elements i and j of the data matrix $X^{m\times p}$ for the attribute column a.

All derivations in the following sections are applicable to nearest-neighbor distance-based methods in general, which includes not only NPDR, but also Relief-based algorithms. Each of these methods uses a distance metric, such as, Eq. 1 to compute neighbors for each instance $i \in \mathcal{I}$. Therefore, our derivations of asymptotic distance distributions are applicable to all methods that compute neighbors in order to weight features. The predictors used by NPDR, however, are the one-dimensional projected distances between two instances $i, j \in \mathcal{I}$ given by Eq. 2. Hence, all asymptotic estimates we derive for Eq. 2 are particularly relevant to NPDR. Since the distance metric given by Eq. 1 is a function of the one-dimensional projection given by Eq. 2, asymptotic estimates derived for Eq. 2 are implicitly relevant to older nearest-neighbor distance-based methods like Relief-based algorithms. We proceed in the following section by applying the Classical Central Limit Theorem and the Delta Method to derive the limit distribution of pairwise distances on any data distribution that is induced by the metric given in Eq. 1.

1.1 Asymptotic normality of pairwise distances

Suppose that $X_{ia}, X_{ja} \overset{iid}{\sim} \mathcal{F}_X(\mu_X, \sigma_X^2)$ for two fixed and distinct instances $i, j \in \mathcal{I}$ 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 $|X_{ia} - X_{ja}|^q = |\mathrm{d}_{ij}(a)|^q$ 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

$$Z_a^q = |\mathrm{d}_{ij}(a)|^q = |X_{ia} - X_{ja}|^q, \quad a \in \mathcal{A}. \tag{3}$$

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

$$\left(D_{ij}^{(q)}\right)^q = \sum_{a\in\mathcal{A}} |\mathrm{d}_{ij}(a)|^q = \sum_{a\in\mathcal{A}} |X_{ia} - X_{ja}|^q = \sum_{a\in\mathcal{A}} Z_a^q \dot{\sim} \mathcal{N}\left(\mu_{z^q} p, \sigma_{z^q}^2 p\right). \tag{4}$$

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 [12] can be applied to show that

$$\begin{aligned} g\left(\left(D_{ij}^{(q)}\right)^q\right) &= g\left(\sum_{a\in\mathcal{A}}^{p} Z_a^q\right) \\ &= \left(\sum_{a\in\mathcal{A}} |X_{ia} - X_{ja}|^q\right)^{1/q} \\ &= D_{ij}^{(q)} \dot{\sim} \mathcal{N}\left(g\left(\mu_{z^q} p\right), \left[g\prime\left(\mu_{z^q} p\right)\right]^2 \sigma_{z^q}^2 p\right) \\ \Rightarrow \quad D_{ij}^{(q)} &\dot{\sim} \mathcal{N}\left(\left(\mu_{z^q} p\right)^{1/q}, \frac{\sigma_{z^q}^2 p}{q^2 \left(\mu_{z^q} p\right)^{2\left(1-\frac{1}{q}\right)}}\right). \end{aligned} \tag{5}$$

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

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2)}\right) &= \sqrt{\mathrm{E}\left[\left(D_{ij}^{(2)}\right)^2\right] - \mathrm{Var}\left(D_{ij}^{(2)}\right)} \\ &= \sqrt{\mu_{z^2} p - \frac{\sigma_{z^2}^2}{4\mu_{z^2}}}. \end{aligned} \tag{6}$$

We estimate rate of convergence to normality for Euclidean ($q = 2$) and Manhattan ($q = 1$) metrics by comparing the distribution of pairwise distances in simulated data to a Gaussian (Fig. 1). We compute the distance between all pairs of instances in simulated datasets of uniformly distributed random data. We simulate data with fixed $m = 100$ instances, and, by varying the number of attributes ($p = 10, 100, 10000$), we observe rapid convergence to Gaussian. For p as low as 10 attributes, Gaussian is a good approximation. The number of attributes in bioinformatics data is typically quite large, at least on the order of 10^3. The Euclidean metric has stronger convergence to a Gaussian than Manhattan. This may be due to Euclidean's use of the square root, which is a common transformation of data in statistics. Normality was assessed using the Shapiro-Wilk test.

To show asymptotic normality of distances, we did not specify whether the data distribution $\mathcal{F}_X$ was discrete or continuous. This is because asymptotic normality is a general phenomenon in high attribute dimension p for any data distribution $\mathcal{F}_X$ satisfying the assumptions we have made. Therefore, Fig. 1 has an analogous representation for discrete data, as well as all other continuous data distributions.

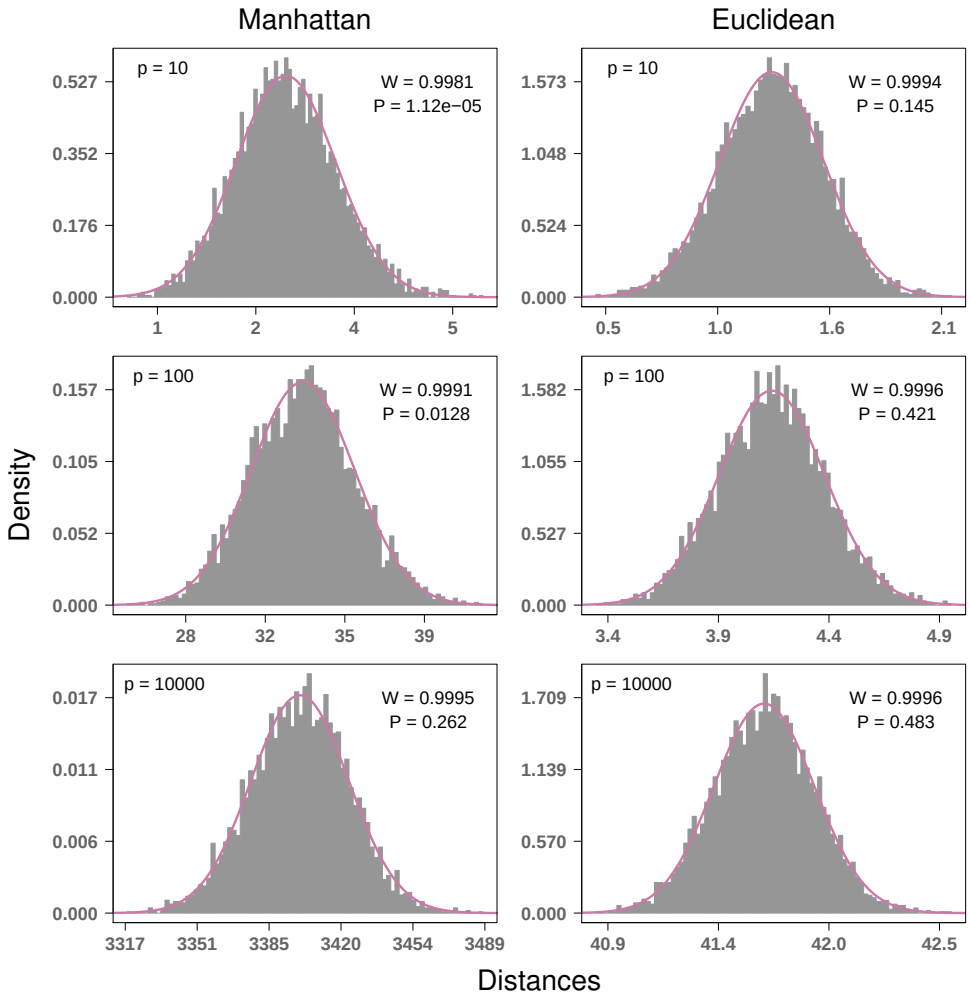

Fig 1. Convergence to Gaussian for Manhattan and Euclidean distances for simulated standard uniform data with $m = 100$ instances and $p = 10, 100$, and 10000 attributes. Convergence to Gaussian occurs rapidly with increasing p, and Gaussian is a good approximation for p as low as 10 attributes. The number of attributes in bioinformatics data is typically much larger, at least on the order of 10^3. The Euclidean metric has stronger convergence to normal than Manhattan. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

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 variance approximately equal to those given in Eq. 5. 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 L_q metric moments for continuous data distributions

In this section, we begin by deriving general formulas for asymptotic means and variances of the L_q distance given by Eq. 1 for standard normal and standard uniform data. With our general formulas for continuous data, we compute moments associated with Manhattan (L_1) and Euclidean (L_2). We then consider the max-min normalized version of the L_q distance, where the magnitude difference given by Eq. 2 is divided by the range of each feature a. Using Extreme Value Theory (EVT), we derive formulas for the moments of feature range in standard normal and standard uniform data. Transitioning into discrete data distributions relevant to GWAS, we derive asymptotic moments for two well known metrics and one new metric. In addition, we derive distance asymptotics for time series correlation-based data, such as, resting-state fMRI.

2.1 Distribution of $|\mathbf{d}_{ij}(a)|^q = |X_{ia} - X_{ja}|^q$

Suppose that $X_{ia}, X_{ja} \overset{iid}{\sim} \mathcal{F}_X(\mu_x, \sigma_x^2)$ and define $Z_a^q = |\mathrm{d}_{ij}(a)|^q = |X_{ia} - X_{ja}|^q$, 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 [13].

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*

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

Elsewhere, $g(y) = 0$.

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

(i) Suppose that $X_{ja} = X_{ia} - (Z_a^q)^{1/q}$. Based on the iid assumption for X_{ia} and X_{ja}, it follows from Thm. 2.1 that the joint density function $g^{(1)}$ of X_{ia} and Z_a^q is given by

$$\begin{aligned} g^{(1)}(x_{ia}, z_a) &= f_X(x_{ia}, x_{ja}) \left| \frac{\partial x_{ja}}{\partial z_a} \right| \\ &= f_X(x_{ia}) f_X(x_{ja}) \left| \frac{-1}{q} \left(z_a^q\right)^{\frac{1}{q}-1} \right| \\ &= \frac{1}{q \left(z_a^q\right)^{1-\frac{1}{q}}} f_X(x_{ia}) f_X\left(x_{ia} - (z_a^q)^{1/q}\right), \quad z_a > 0 \end{aligned} \tag{7}$$

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

$$\begin{aligned} f_{Z_a^q}^{(1)}(z_a^q) &= \int_{-\infty}^{\infty} g^{(1)}(x_{ia}, z_a^q) \mathrm{d}x_{ia} \\ &= \frac{1}{q \left(z_a^q\right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) f_X\left(x_{ia} - (z_a^q)^{1/q}\right) \mathrm{d}x_{ia}, \quad z_a > 0. \end{aligned} \tag{8}$$

(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

$$\begin{aligned} g^{(2)}(x_{ia}, z_a) &= f_X(x_{ia}, x_{ja}) \left| \frac{\partial x_{ja}}{\partial z_a} \right| \\ &= f_X(x_{ia}) f_X(x_{ja}) \left| \frac{1}{q} \left(z_a^q\right)^{\frac{1}{q}-1} \right| \\ &= \frac{1}{q \left(z_a^q\right)^{1-\frac{1}{q}}} f_X(x_{ia}) f_X\left(x_{ia} - (z_a^q)^{1/q}\right), \quad z_a > 0. \end{aligned} \tag{9}$$

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

$$\begin{aligned} f_{Z_a^q}^{(2)}(z_a^q) &= \int_{-\infty}^{\infty} g^{(2)}(x_{ia}, z_a^q) \mathrm{d}x_{ia} \\ &= \frac{1}{q \left(z_a^q\right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) f_X\left(x_{ia} + (z_a^q)^{1/q}\right) \mathrm{d}x_{ia}, \quad z_a > 0. \end{aligned} \tag{10}$$

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

$$E^{(1)} = \left\{ |X_{ia} - X_{ja}|^q \leq z_a^q | X_{ja} = X_{ia} - (Z_a^q)^{1/q} \right\} \tag{11}$$

and

$$E^{(2)} = \left\{ |X_{ia} - X_{ja}|^q \leq z_a^q | X_{ja} = X_{ia} + (Z_a^q)^{1/q} \right\}. \tag{12}$$

Then it follows from fundamental rules of probability that

$$\begin{aligned} F_{Z_a^q}(z_a^q) &= \mathrm{P}\left[Z_a^q \leq z_a^q\right] \\ &= \mathrm{P}\left[|X_{ia} - X_{ja}|^q \leq z_a^q\right] \\ &= \mathrm{P}\left[E^{(1)} \cup E^{(2)}\right] \\ &= \mathrm{P}\left[E^{(1)}\right] + \mathrm{P}\left[E^{(2)}\right] - \mathrm{P}\left[E^{(1)} \cap E^{(2)}\right] \\ &= \mathrm{P}\left[E^{(1)}\right] + \mathrm{P}\left[E^{(2)}\right] \\ &= \int_{-\infty}^{z_a^q} f_{Z_a^q}^{(1)}(t) \mathrm{d}t + \int_{-\infty}^{z_a^q} f_{Z_a^q}^{(2)}(t) \mathrm{d}t \\ &= \int_{-\infty}^{z_a^q} \left(f_{Z_a^q}^{(1)}(t) + f_{Z_a^q}^{(2)}(t) \right) \mathrm{d}t \\ &= \frac{1}{q \left(z_a^q\right)^{1-\frac{1}{q}}} \int_{-\infty}^{z_a^q} \left(\int_{-\infty}^{\infty} f_X(x_{ia}) \left[f_X(x_{ia} - t) + f_X(x_{ia} + t) \right] \mathrm{d}x_{ia} \right) \mathrm{d}t, \quad z_a > 0. \end{aligned} \tag{13}$$

It follows directly from the result in Eq. 13 that the density function of the random variable Z_a^q is given by

$$\begin{aligned} f_{Z_a^q}(z_a^q) &= \frac{\partial}{\partial z_a^q} F_{Z_a^q}(z_a^q) \\ &= \frac{1}{q \left(z_a^q\right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[f_X\left(x_{ia} - (z_a^q)^{1/q}\right) + f_X\left(x_{ia} + (z_a^q)^{1/q}\right) \right] \mathrm{d}x_{ia}, \end{aligned} \tag{14}$$

where $z_a > 0$.

Using Eq. 14, we can compute the mean and variance of the random variable Z_a^q as

$$\mu_{z_a^q} = \int_{-\infty}^{\infty} z_a^q f_{Z_a^q}(z_a^q) \mathrm{d}z_a^q \tag{15}$$

and

$$\sigma_{z_a^q}^2 = \int_{-\infty}^{\infty} \left(z_a^q\right)^2 f_{Z_a^q}(z_a^q) \mathrm{d}z_a^q - \mu_{z_a^q}^2. \tag{16}$$

It follows immediately from Eqs. 15 and 16 and the Classical Central Limit Theorem (CCLT) that

$$\left(D_{ij}^{(q)}\right)^q = \sum_{a\in\mathcal{A}} Z_a^q = \sum_{a\in\mathcal{A}} |X_{ia} - X_{ja}|^q \dot{\sim} \mathcal{N}\left(\mu_{z^q} p, \sigma_{z^q}^2 p\right). \tag{17}$$

Applying the result given in Eq. 5, the distribution of $D_{ij}^{(q)}$ is given by

$$D_{ij}^{(q)} \dot{\sim} \mathcal{N}\left(\left(\mu_{z_a^q} p\right)^{1/q}, \frac{\sigma_{z_a^q}^2 p}{q^2 \left(\mu_{z_a^q} p\right)^{2\left(1-\frac{1}{q}\right)}}\right), \quad \mu_{z_a^q} > 0 \tag{18}$$

with improved estimate of the mean for $q = 2$ given by Eq. 6.

2.1.1 Standard normal data

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

$$f_X(x_{ia}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_{ia}^2}, \tag{19}$$

$$f_X\left(x_{ia} - (z_a^q)^{1/q}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x_{ia} - (z_a^q)^{1/q}\right)^2}, \quad z_a > 0, \text{ and} \tag{20}$$

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

Substituting the results given by Eqs. 19-21 into Eq. 14 and completing the square on x_{ia} in the exponents, we have

$$\begin{aligned}
f_{Z_a^q}(z_a^q) &= \frac{1}{2q\pi \left(z_a^q\right)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \Bigg(e^{-\frac{1}{2}\left[\sqrt{2}x_{ia} - \frac{\sqrt{2}}{2}(z_a^q)^{1/q}\right]^2} \\
&\qquad + e^{-\frac{1}{2}\left[\sqrt{2}x_{ia} + \frac{\sqrt{2}}{2}(z_a^q)^{1/q}\right]^2} \Bigg) \mathrm{d}x_{ia} \\
&= \frac{1}{2q\sqrt{\pi}\left(z_a^q\right)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left(e^{-\frac{1}{2}u^2} + e^{-\frac{1}{2}u^2}\right) \mathrm{d}u \\
&= \frac{1}{2q\sqrt{\pi}\left(z_a^q\right)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} (1+1) \\
&= \frac{1}{q\sqrt{\pi}} \left(z_a^q\right)^{\frac{1}{q}-1} e^{-\frac{1}{4}(z_a^q)^{2/q}} \\
&= \frac{\frac{2}{q}}{(2^q)^{1/q}\,\Gamma\left(\frac{\frac{1}{q}}{\frac{2}{q}}\right)} \left(z_a^q\right)^{\frac{1}{q}-1} e^{-\left(\frac{z_a^q}{2^q}\right)^{2/q}}.
\end{aligned} \tag{22}$$

The density function given by Eq. 22 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

$$\begin{aligned} \mu_{z_a^q} &= \frac{c\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \\ &= \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} \end{aligned} \tag{23}$$

and

$$\begin{aligned} \sigma^2_{z_a^q} &= c^2\left[\frac{\Gamma\left(\frac{d+2}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} - \left(\frac{\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)}\right)^2\right] \\ &= 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]. \end{aligned} \tag{24}$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 23 and 24 allow the p-dimensional mean and variance of the $D_{ij}^{(q)}$ distribution to be computed directly as

$$\mu_{\left(D_{ij}^{(q)}\right)^q} = \mathrm{E}\left[\left(D_{ij}^{(q)}\right)^q\right] = \mathrm{E}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) = \sum_{a\in\mathcal{A}} \mathrm{E}\left(Z_a^q\right) = \sum_{a\in\mathcal{A}} \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} = \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}p \tag{25}$$

and

$$\begin{aligned} \sigma^2_{\left(D_{ij}^{(q)}\right)^q} = \mathrm{Var}\left[\left(D_{ij}^{(q)}\right)^q\right] &= \mathrm{Var}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) \\ &= \sum_{a\in\mathcal{A}} \mathrm{Var}\left(Z_a^q\right) \\ &= \sum_{a\in\mathcal{A}} 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right] \\ &= 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]p. \end{aligned} \tag{26}$$

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

$$\mathcal{N}\left(\left(2^q\frac{\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/q}, \frac{4^qp}{q^2\left(\frac{2^q\Gamma\left(\frac{1}{2}q+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]\right). \tag{27}$$

We provide a summary table of the moment estimates (Eq. 27) for the L_q metric on standard normal data (Tab. 1). The summary is organized by data type, type of statistic (mean or variance), and corresponding asymptotic formula. In the next section, we derive the L_q distance distribution on standard uniform data in a similar fashion.

2.1.2 Standard uniform data

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

$$f_X(x_{ia}) = 1, \quad 0 \leq x_{ia} \leq 1 \tag{28}$$

$$f_X\left(x_{ia} - (z_a^q)^{1/q}\right) = 1, \quad 0 \leq x_{ia} - (z_a^q)^{1/q} \leq 1, \text{ and} \tag{29}$$

$$f_X\left(x_{ia} + (z_a^q)^{1/q}\right) = 1, \quad 0 \leq x_{ia} + (z_a^q)^{1/q} \leq 1. \tag{30}$$

Substituting the results given by Eqs. 28-30 into Eq. 14, we have

$$\begin{aligned}
f_{Z_a^q}(z_a^q) &= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[f_X\left(x_{ia} - (z_a^q)^{1/q}\right) + f_X\left(x_{ia} + (z_a^q)^{1/q}\right)\right] \mathrm{d}x_{ia}, \\
&\qquad\qquad 0 < z_a \leq 1 \\
&= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}} \int_0^1 \left[f_X(x_{ia} - (z_a^q) + f_X\left(x_{ia} + (z_a^q)^{1/q}\right)\right] \mathrm{d}x_{ia}, \quad 0 < z_a \leq 1 \\
&= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}} \int_{(z_a^q)}^1 1\mathrm{d}x_{ia} + \int_0^{1-(z_a^q)} 1\mathrm{d}x_{ia}, \quad 0 < z_a \leq 1 \\
&= \frac{1}{q\left(z_a^q\right)^{1-\frac{1}{q}}} \left[(1 - (z_a^q)) + (1 - (z_a^q))\right], \quad 0 < z_a \leq 1 \\
&= \frac{1}{q} \cdot 2\left(z_a^q\right)^{\frac{1}{q}-1} \left[1 - (z_a^q)^{1/q}\right]^{2-1}, \quad 0 < z_a \leq 1.
\end{aligned} \tag{31}$$

The density given by Eq. 31 is a Kumaraswamy density with parameters $b = \frac{1}{q}$ and $c = 2$ with moment generating function (MGF) given by

$$\begin{aligned}
M_n &= \frac{c\Gamma\left(1 + \frac{n}{b}\right)\Gamma(c)}{\Gamma\left(1 + c + \frac{n}{b}\right)} \\
&= \frac{2}{(nq+2)(nq+1)}.
\end{aligned} \tag{32}$$

Using the MGF given by Eq. 32, the mean and variance of Z_a^q are computed as

$$\mu_{z_a^q} = \frac{2}{(q+2)(q+1)} \tag{33}$$

and

$$\sigma^2_{z_a^q} = \frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2. \tag{34}$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 35 and 36 allow the p-dimensional mean and variance of the $\left(D_{ij}^{(q)}\right)^q$ distribution to be computed directly as

$$\begin{aligned}
\mu_{\left(D_{ij}^{(q)}\right)^q} = \mathrm{E}\left[\left(D_{ij}^{(q)}\right)^q\right] &= \mathrm{E}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) \\
&= \sum_{a\in\mathcal{A}} \mathrm{E}(Z_a^q) \\
&= \sum_{a\in\mathcal{A}} \frac{2}{(q+2)(q+1)} \\
&= \frac{2p}{(q+2)(q+1)}
\end{aligned} \tag{35}$$

and

$$\begin{aligned}\sigma^2_{\left(D_{ij}^{(q)}\right)^q} = \operatorname{Var}\left[\left(D_{ij}^{(q)}\right)^q\right] &= \operatorname{Var}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) \\ &= \sum_{a\in\mathcal{A}} \operatorname{Var}\left(Z_a^q\right) \\ &= \sum_{a\in\mathcal{A}} \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right] \\ &= \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right] p.\end{aligned} \tag{36}$$

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

$$\mathcal{N}\left(\left(\frac{2p}{(q+2)(q+1)}\right)^{1/q}, \frac{p}{q^2\left(\frac{2p}{(q+2)(q+1)}\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right]\right). \tag{37}$$

We provide a summary table of the moment estimates (Eq. 37) for the L_q metric on standard uniform data (Tab. 1). The summary is organized by data type, type of statistic (mean or variance), and corresponding asymptotic formula. In the next section, we use our general L_q distance distribution derivations to provide Manhattan ($q = 1$) and Euclidean ($q = 2$) asymptotic moments on both standard normal and standard uniform data. These are the most commonly applied metrics in the context of nearest-neighbor feature selection, so they are of particular interest.

2.2 Manhattan (L_1)

With our general formulas for the asymptotic mean and variance (Eqs. 27 and 37) for any value of $q \in \mathbb{N}$, we can simply substitute a particular value of q in order to determine the asymptotic distribution of the corresponding distance L_q metric. We demonstrate this with the example of the Manhattan metric (L_1) for standard normal and standard uniform data (Eq. 1, $q = 1$).

2.2.1 Standard normal data

Substituting $q = 1$ into the asymptotic formula for the mean L_q distance (Eq. 27), we have the following for expected L_1 distance between two independently sample instances $i, j \in \mathcal{I}$ in standard normal data

$$\begin{aligned}\mathrm{E}\left(D_{ij}^{(1)}\right) &= \left(2\frac{\Gamma\left(\frac{1+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/1} \\ &= \frac{2p}{\sqrt{\pi}}.\end{aligned} \tag{38}$$

We see in Eq. 38 that $D_{ij}^{(1)} \sim p$ in the limit, which implies that this distance is unbounded as feature dimension p increases.

Substituting $q = 1$ into the formula for the asymptotic variance of $D_{ij}^{(1)}$ (Eq. 27) leads to the following

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(1)}\right) &= \frac{4^1 p}{1^2\left(\frac{2^1\Gamma\left(\frac{1}{2}(1)+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{1}\right)}}\left[\frac{\Gamma\left(1+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}(1)+\frac{1}{2}\right)}{\pi}\right] \\ &= \frac{2(\pi-2)p}{\pi}. \end{aligned} \tag{39}$$

Similar to the mean (Eq. 38), the limiting variance of $D_{ij}^{(1)}$ (Eq. 39) grows on the order of feature dimension p, which implies that points become more dispersed as the dimension increases. The moment estimates given in this section (Eqs. 38 and 39) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 2).

2.2.2 Standard uniform data

Substituting $q = 1$ into the asymptotic formula of the mean (Eq. 37), we have the following for the expected L_1 distance between two independently sampled instances $i, j \in \mathcal{I}$ in standard uniform data

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(1)}\right) &= \left(\frac{2p}{(1+2)(1+1)}\right)^{1/1} \\ &= \frac{p}{3}. \end{aligned} \tag{40}$$

Once again, we see that the mean of $D_{ij}^{(1)}$ (Eq. 40) grows on the order of p just as in the case of standard normal data.

Substituting $q = 1$ into the formula of the asymptotic variance of $D_{ij}^{(1)}$ (Eq. 37) leads to the following

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(1)}\right) &= \frac{p}{1^2\left(\frac{2p}{(1+2)(1+1)}\right)^{2\left(1-\frac{1}{1}\right)}}\left[\frac{1}{(1+1)(2(1)+1)} - \left(\frac{2}{(1+2)(1+1)}\right)^2\right] \\ &= \frac{p}{18}. \end{aligned} \tag{41}$$

As in the case of the L_1 metric on standard normal data, we have a variance (Eq. 41) that grows on the order of p. The distances between points in high-dimensional uniform data become more widely dispersed with this metric. The moment estimates given in this section (Eqs. 40 and 41) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 2).

2.2.3 Distribution of one-dimensional projection of pairwise distance onto an attribute

In nearest-neighbor distance-based feature selection like NPDR and Relief-based algorithms, the one-dimensional projection of the pairwise distance onto an attribute (Eq. 2) is particularly fundamental to feature quality. For instance, this distance projection is the predictor used to determine beta coefficients for features in NPDR. In particular, understanding distributional properties of the projected distances is necessary for defining P values for NPDR. Up to this point, the distribution of NPDR beta coefficients under the null hypothesis is an unsolved problem. It is crucial to solve this problem in order

to have realistic P values from computed beta coefficients in NPDR. In this section, we summarize the exact distribution of the one-dimensional projected distance onto an attribute $a \in \mathcal{A}$. These results apply to continuous data, such as, gene expression.

In previous sections, we derived the exact density function (Eq. 14) and moments (Eqs. 15 and 16) for the distribution of $Z_a^q = |X_{ia} - X_{ja}|^q$. We then derived the exact density (Eq. 22) and moments (Eqs. 23 and 24) for standard normal data. Analogously, we formulated the exact density (Eq. 31) and moments (Eqs. 33 and 34) for standard uniform data. From these exact densities and moments, we simply substitute $q = 1$ to define the distribution of the one-dimensional projected distance onto an attribute $a \in \mathcal{A}$.

Assuming data is standard normal, we substitute $q = 1$ into Eq. 22 to arrive at the following density function

$$\begin{aligned} f_{Z_a^1}\left(z_a^1\right) &= \frac{\frac{2}{1}}{(2^1)^{1/1}\,\Gamma\left(\frac{\frac{1}{1}}{\frac{2}{1}}\right)} \left(z_a^1\right)^{1/1-1} e^{-\left(\frac{z_a^1}{\frac{2}{1}}\right)^{2/1}}, \quad z_a > 0 \\ &= \frac{1}{\sqrt{\pi}} z_a e^{-\frac{1}{4}z_a^2}, \quad z_a > 0. \end{aligned} \tag{42}$$

The mean corresponding to this Generalized Gamma density is computed by substituting $q = 1$ into Eq. 23. This result is given by

$$\begin{aligned} \mu_{Z_a^1} &= \frac{2^1 \Gamma\left(\frac{1+1}{2}\right)}{\sqrt{\pi}} \\ &= \frac{2}{\sqrt{\pi}}. \end{aligned} \tag{43}$$

Substituting $q = 1$ into Eq. 24 for the variance, we have the following

$$\begin{aligned} \sigma^2_{Z_a^1} &= 4^1 \left[\frac{\Gamma\left(1+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}\cdot 1 + \frac{1}{2}\right)}{\pi}\right] \\ &= \frac{2(\pi-2)}{\pi}. \end{aligned} \tag{44}$$

These last few results (Eqs. 42-44) provide us with the distribution for NPDR predictors when the data is from the standard normal distribution.

If we have standard uniform data, we substitute $q = 1$ into Eq. 31 to obtain the following density function

$$\begin{aligned} f_{Z_a^1} &= \frac{1}{1} \cdot 2\left(z_a^1\right)^{1/1-1} \left[1 - \left(z_a^1\right)^{1/1}\right]^{2-1}, \quad 0 < z_a \leq 1 \\ &= 2z_a(1-z_a), \quad 0 < z_a \leq 1. \end{aligned} \tag{45}$$

The mean corresponding to this Kumaraswamy density is computed by substituting $q = 1$ into Eq. 33. After substitution, we have the following result

$$\begin{aligned} \mu_{Z_a^1} &= \frac{2}{(1+2)(1+1)} \\ &= \frac{1}{3}. \end{aligned} \tag{46}$$

Substituting $q = 1$ into Eq. 34 for the variance, we have the following

$$\begin{aligned} \sigma^2_{Z_a^1} &= \frac{1}{(1+1)(2\cdot 1+1)} - \left(\frac{2}{(1+2)(1+1)}\right)^2 \\ &= \frac{1}{18}. \end{aligned} \tag{47}$$

In the event that the data distribution is standard uniform, Eqs. 45-47 provide the distribution for NPDR predictors. The means (Eqs. 43 and 46) and variances (Eqs.44 and 47) come from the exact distribution of pairwise distances with respect to a single attribute $a \in \mathcal{A}$. This is the distribution of the so-called "projection" of the pairwise distance onto a single attribute to which we have been referring, which is a direct implication from our more general derivations. In a similar manner, one can substitute any value of $q \geq 2$ into Eqs. 22 and 31 to derive the associated density of $Z_a^q = |X_{ia} - X_{ja}|^q$ for the given data type.

2.3 Euclidean (L_2)

Moment estimates for the Euclidean metric are obtained by substituting $q = 2$ into the asymptotic moment formulas for standard normal data (Eq. 27) and standard uniform data (Eq. 37). As in the case of the Manhattan metric in the previous sections, we initially proceed by deriving Euclidean distance moments in standard normal data.

2.3.1 Standard normal data

Substituting $q = 2$ into the asymptotic formula of the mean (Eq. 27), we have the following for expected L_2 distance between two independently sampled instances $i, j \in \mathcal{I}$ in standard normal data

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2)}\right) &= \left(2\frac{\Gamma\left(\frac{2+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/2} \\ &= \sqrt{2p}. \end{aligned} \tag{48}$$

In the case of L_2 on standard normal data, we see that the mean of $D_{ij}^{(2)}$ (Eq. 48) grows on the order of $\sqrt{p}$. Hence, the Euclidean distance does not increase as quickly as the Manhattan distance on standard normal data.

Substituting $q = 2$ into the formula for the asymptotic variance of $D_{ij}^{(2)}$ (Eq. 27) leads to the following

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(2)}\right) &= \frac{4^2 p}{2^2\left(\frac{2^2\Gamma\left(\frac{1}{2}(2)+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{2}\right)}}\left[\frac{\Gamma\left(2+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}(2)+\frac{1}{2}\right)}{\pi}\right] \\ &= 1. \end{aligned} \tag{49}$$

Surprisingly, the asymptotic variance (Eq. 49) is just 1. Regardless of data dimensions m and p, the variance of Euclidean distances on standard normal data tends to 1. Therefore, most instances are contained within a ball of radius 1 about the mean in high feature dimension p. This means that the Euclidean distance distribution on standard normal data is simply a horizontal shift to the right of the standard normal distribution.

For the case in which the number of attributes p is small, we have an improved estimate of the mean Eq. 6). The lower dimensional estimate of the mean is given by

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

For high dimensional data sets like gene expression [14, 15], which typically contain thousands of genes (or features), it is clear that the magnitude of p will be sufficient to

use the standard asymptotic estimate (Eq. 48) since $\sqrt{2p} \approx \sqrt{2p-1}$ in that case. The moment estimates given in this section (Eqs. 50 and 49) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 2).

2.3.2 Standard uniform data

Substituting $q = 2$ into the asymptotic formula of the mean (Eq. 37), we have the following for expected L_2 distance between two independently sampled instances $i, j \in \mathcal{I}$ in standard uniform data

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

As in the case of standard normal data, the expected value of $D_{ij}^{(2)}$ (Eq. 51) grows on the order of $\sqrt{p}$.

Substituting $q = 2$ into the formula for the asymptotic variance of $D_{ij}^{(2)}$ (Eq. 37) leads to the following

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(2)}\right) &= \frac{p}{2^2\left(\frac{2p}{(2+2)(2+1)}\right)^{2\left(1-\frac{1}{2}\right)}}\left[\frac{1}{(2+1)(2(2)+1)} - \left(\frac{2}{(2+2)(2+1)}\right)^2\right] \\ &= \frac{7}{120}. \end{aligned} \tag{52}$$

Once again, the variance of Euclidean distance surprisingly approaches a constant.

For the case in which the number of attributes p is small, we have an improved estimate of the mean (Eq. 6). The lower dimensional estimate of the mean is given by

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2)}\right) &= \left(\frac{2p}{(2+2)(2+1)} - \frac{7}{120}\right)^{1/2} \\ &= \sqrt{\frac{p}{6} - \frac{7}{120}}. \end{aligned} \tag{53}$$

The moment estimates given in this section (Eqs. 53 and 52) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 2). This concludes our analysis with continuous data distributions and the standard L_q metric. In the next section, we will use extreme value theory to derive the distribution of the sample maximum and minimum for standard normal and standard uniform data. This will lead us to asymptotics for the max-min normalized L_q metric used frequently in Relief-based algorithms [1,3] for scoring features.

2.4 Distribution of max-min normalized L_q metric

For Relief-based methods [1,3], the standard numeric diff metric is given by

$$\mathrm{d}_{ij}^{\mathrm{num}}(a) = \mathrm{diff}(a,(i,j)) = \frac{|X_{ia} - X_{ja}|}{\max(a) - \min(a)}, \tag{54}$$

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

The pairwise distance using this max-min normalized diff metric is then computed as

$$
\begin{aligned}
D_{ij}^{(q*)} &= \left(\sum_{a \in \mathcal{A}} |\mathrm{d}_{ij}(a)|^q \right)^{1/q} \\
&= \left(\sum_{a \in \mathcal{A}} \left(\frac{|X_{ia} - X_{ja}|}{\max(a) - \min(a)} \right)^q \right)^{1/q}.
\end{aligned}
\tag{55}
$$

In order to determine moments of asymptotic distance distributions induced by Eq. 54, we must first derive the asymptotic extreme value distributions of the attribute maximum and minimum. Although the exact distribution of the maximum or minimum requires an assumption about the data distribution, the Fisher-Tippett-Gnedenko Theorem is an important result that allows one to generally categorize the extreme value distribution for a collection of independent and identically distributed random variables into one of three distributional families. This theorem does not, however, tell us the exact distribution of the maximum that we require in order to determine asymptotic results for the max-min normalized distance (Eq. 55). We mention this theorem simply to provide some background on convergence of extreme values. Before stating the theorem, we first need the following definition

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

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

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

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

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

The three distribution families given in Thm. 2.2 are actually special cases of the Generalized Extreme Value Distribution. In the context of extreme values, Thm. 2.2 is analogous to the Central Limit Theorem for the distribution of sample mean. Although we will not explicitly invoke this theorem, it does tell us something very important about the asymptotic behavior of sample extremes under certain necessary conditions. For illustration of this general phenomenon of sample extremes, we derive the distribution of the maximum for standard normal data to show that the limiting distribution is in the Gumbel family, which is a well known result. In the case of standard uniform data, we will derive the distribution of the maximum and minimum directly. Regardless of data

type, the distribution of the sample maximum can be derived as follows

$$\begin{aligned} \mathrm{P}[X_a^{\max} \leq x] &= \mathrm{P}\left[\max_{k\in\mathcal{I}}\{X_{ka}\} \leq x\right] \\ &= \mathrm{P}[X_{1a} \leq x, X_{2a} \leq x, \ldots, X_{ma} \leq x] \\ &= \prod_{k=1}^{m} \mathrm{P}[X_{ka} \leq x] \\ &= \prod_{k=1} F_X(x) \\ &= [F_X(x)]^m. \end{aligned} \tag{56}$$

Using more precise notation, the distribution function of the sample maximum in standard normal data is

$$F_{\max}(x) = [F_X(x)]^m, \tag{57}$$

where m is the size of the sample from which the maximum is derived and F_X is the distribution function corresponding to the data sample. This means that the distribution of the sample maximum relies only on the distribution function of the data from which extremes are drawn F_X and the size of the sample m.

Differentiating the distribution function (Eq. 57) gives us the following density function for the distribution of the maximum

$$\begin{aligned} f_{\max}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{\max}(x) \\ &= \frac{\mathrm{d}}{\mathrm{d}x} [F_X(x)]^m \\ &= m[F_X(x)]^{m-1} f_X(x), \end{aligned} \tag{58}$$

where m is the size of the sample from which the maximum is derived, F_X is the distribution function corresponding to the data sample, and f_X is the density function corresponding to the data sample. Similar to the distribution function for the sample maximum (Eq. 57), the density function (Eq 58) relies only on the distribution and density function of the data from which extremes are derived.

The distribution of the sample minimum, $X_a^{\min}$, can be derived as follows

$$\begin{aligned} \mathrm{P}[X_a^{\min} \leq x] &= 1 - \mathrm{P}[X_a^{\min} \geq x] \\ &= 1 - \mathrm{P}\left[\min_{k\in\mathcal{I}}\{X_{ka}\} \geq x\right] \\ &= 1 - \mathrm{P}[X_{1a} \geq x, X_{2a} \geq x, \ldots, X_{ma} \geq x] \\ &= 1 - \prod_{k=1}^{m} \mathrm{P}[X_{ka} \geq x] \\ &= 1 - \left[\mathrm{P}[X_{1a} \geq x]\right]^m \\ &= 1 - \left[1 - \mathrm{P}[X_{1a} \leq x]\right]^m \\ &= 1 - \left[1 - F_X(x)\right]^m, \end{aligned} \tag{59}$$

where m is the size of the sample from which the maximum is derived and F_X is the distribution function corresponding to the data sample. Therefore, the distribution of sample minimum also relies only on the distribution function of the data from which extremes are derived.

With more precise notation, we have the following expression for the distribution function of the minimum

$$F_{\min}(x) = 1 - [1 - F_X(x)]^m. \tag{60}$$

where m is the size of the sample from which the minimum is derived and F_X is the distribution function corresponding to the data sample.

Differentiating the distribution function (Eq. 60) gives us the following density function for the distribution of sample minimum

$$\begin{aligned} f_{\min}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{\min}(x) \\ &= \frac{\mathrm{d}}{\mathrm{d}x} \left(1 - [1 - F_X(x)]^m\right) \\ &= m \left[1 - F_X(x)\right]^{m-1} f_X(x), \end{aligned} \tag{61}$$

where m is the size of the sample from which the minimum is derived, F_X is the distribution function corresponding to the data sample, and f_X is the density function corresponding to the data sample. As in the case of the density function for sample maximum (Eq. 58), the density function for sample minimum relies only on the distribution F_X and density f_X functions of the data from which extremes are derived and the sample size m.

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

$$\begin{aligned} \mu^{(1)}_{\max}(m) = \mathrm{E}(X_a^{\max}) &= \int_{-\infty}^{\infty} x f_{\max}(x) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} x \left(m[F_X(x)]^{m-1} f_X(x)\right) \mathrm{d}x \\ &= m \int_{-\infty}^{\infty} x f_X(x) [F_X(x)]^{m-1} \mathrm{d}x, \end{aligned} \tag{62}$$

where m is the sample size, F_X is the distribution function, and f_X is the density function of the data from which the maximum is derived.

The second raw moment of the distribution of sample maximum is derived similarly as follows

$$\begin{aligned} \mu^{(2)}_{\max}(m) = \mathrm{E}[(X_a^{\max})^2] &= \int_{-\infty}^{\infty} x^2 f_{\max}(x) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2 \left(m[F_X(x)]^{m-1} f_X(x)\right) \mathrm{d}x \\ &= m \int_{-\infty}^{\infty} x^2 f_X(x) [F_X(x)]^{m-1} \mathrm{d}x \end{aligned} \tag{63}$$

where m is the sample size, F_X is the distribution function, and f_X is the density function of the data from which the maximum is derived.

Using the first (Eq. 62) and second (Eq. 63) raw moments of the distribution of sample maximum, the variance is given by

$$\sigma^2_{\max}(m) = \mu^{(2)}_{\max}(m) - \left[\mu^{(1)}_{\max}(m)\right]^2, \tag{64}$$

where m is the sample size of the data from which the maximum is derived and $\mu^{(1)}_{\max}(m)$ and $\mu^{(2)}_{\max}$ are the first and second raw moments, respectively, of the distribution of sample maximum.

Moving on to the distribution of sample minimum, the first raw moment is given by

the following

$$\begin{aligned}\mu^{(1)}_{\min}(m) = \mathrm{E}(X^{\min}_a) &= \int_{-\infty}^{\infty} x f_{\min}(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x\left(m[F_X(x)]^{m-1} f_X(x)\right)\mathrm{d}x \\ &= m\int_{-\infty}^{\infty} x f_X(x)[F_X(x)]^{m-1}\mathrm{d}x,\end{aligned} \tag{65}$$

where m is the sample size, F_X is the distribution function, and f_X is the density function of the data from which the minimum is derived.

Similarly, the second raw moment of the distribution of sample minimum is given by the following

$$\begin{aligned}\mu^{(2)}_{\min}(m) = \mathrm{E}[(X^{\min}_a)^2] &= \int_{-\infty}^{\infty} x^2 f_{\min}(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2\left(m[F_X(x)]^{m-1} f_X(x)\right)\mathrm{d}x \\ &= m\int_{-\infty}^{\infty} x^2 f_X(x)[F_X(x)]^{m-1}\mathrm{d}x,\end{aligned} \tag{66}$$

where m is the sample size, F_X is the distribution function, and f_X is the density function of the data from which the minimum is derived.

Using the first (Eq. 65) and second (Eq. 66) raw moments of the distribution of sample minimum, the variance is given by

$$\sigma^2_{\min}(m) = \mu^{(2)}_{\min}(m) - \left[\mu^{(1)}_{\min}(m)\right]^2, \tag{67}$$

where m is the sample size of the data from which the maximum is derived and $\mu^{(1)}_{\min}(m)$ and $\mu^{(2)}_{\min}$ are the first and second raw moments, respectively, of the distribution of sample maximum.

Using the expected attribute maximum (Eq. 62) and minimum (Eq. 65) for sample size m, the following expected attribute range results from linearity of the expectation operator

$$\begin{aligned}\mathrm{E}(X^{\max}_a - X^{\min}_a) &= \mathrm{E}(X^{\max}_a) - \mathrm{E}(X^{\min}_a) \\ &= \mu^{(1)}_{\max}(m) - \mu^{(1)}_{\min}(m).\end{aligned} \tag{68}$$

where $\mu^{(1)}_{\max}(m)$ is the expected sample maximum (Eq. 62) and $\mu^{(1)}_{\min}(m)$ is the expected sample minimum.

For a data distribution that has zero skewness and support that is symmetric about 0, the expected attribute range (Eq. 68) can be simplified to the following expression

$$\mathrm{E}(X^{\max}_a - X^{\min}_a) = 2\mu^{(1)}_{\max}(m), \tag{69}$$

where m is the size of the sample from which the maximum is derived. Hence, the expected attribute range is simply twice the expected attribute maximum (Eq. 62). This result naturally applies to standard normal data, which is symmetric about its mean at 0 and without any skewness.

For large samples ($m >> 1$) from an exponential type distribution that has infinite support and all moments, the covariance between the sample maximum and minimum is

approximately zero [16]. In this case, the variance of the attribute range of a sample of size m is given by the following

$$\begin{aligned} \mathrm{Var}(X_a^{\max} - X_a^{\min}) &\approx \mathrm{Var}(X_a^{\max}) + \mathrm{Var}(X_a^{\min}) \\ &= \sigma^2_{\max}(m) + \sigma^2_{\min}(m). \end{aligned} \tag{70}$$

Under the assumption of zero skewness, infinite support that is symmetric about 0, sufficiently large sample size m, and distribution of an exponential type with all moments, the variance of attribute range (Eq. 70) simplifies to become the following

$$\begin{aligned} \mathrm{Var}(X_a^{\max} - X_a^{\min}) &= 2\mathrm{Var}(X_a^{\max}) \\ &= 2\sigma^2_{\max}. \end{aligned} \tag{71}$$

Let $\mu_{D_{ij}^{(q)}}$ and $\sigma^2_{D_{ij}^{(q)}}$ (Eq. 18) denote the mean and variance of the standard L_q distance metric (Eq. 1). Then the expected value of the max-min normalized distance (Eq. 55) distribution is given by the following

$$\begin{aligned} \mu_{D_{ij}^{(q*)}} &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}}\right)^q\right)^{1/q}\right] \\ &\approx \frac{1}{\mathrm{E}(X_a^{\max} - X_a^{\min})}\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}|X_{ia} - X_{ja}|^q\right)^{1/q}\right] \\ &= \frac{\mu_{D_{ij}^{(q)}}}{\mathrm{E}(X_a^{\max}) - \mathrm{E}(X_a^{\min})} \\ &= \frac{\mu_{D_{ij}^{(q)}}}{\mu^{(1)}_{\max}(m) - \mu^{(1)}_{\min}(m)}, \end{aligned} \tag{72}$$

where m is the size of the sample from which extremes are derived, $\mu^{(1)}_{\max}(m)$ is the expected value of the sample maximum (Eq. 62), and $\mu^{(1)}_{\min}$ is the expected value of the sample minimum.

The variance of the max-min normalized distance (Eq. 55) distribution is given by

the following

$$
\begin{aligned}
\sigma^2_{D_{ij}^{(q*)}} &= \mathrm{Var}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia}-X_{ja}|}{X_a^{\max}-X_a^{\min}}\right)^q\right)^{1/q}\right] \\
&= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia}-X_{ja}|}{X_a^{\max}-X_a^{\min}}\right)^q\right)^{2/q}\right] - \left(\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}\left(\frac{|X_{ia}-X_{ja}|}{X_a^{\max}-X_a^{\min}}\right)^q\right)^{1/q}\right]\right)^2 \\
&\approx \frac{\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}|X_{ia}-X_{ja}|^q\right)^{2/q}\right]}{\mathrm{E}[(X_a^{\max}-X_a^{\min})^2]} - \frac{\left(\mathrm{E}\left[\left(\sum_{a\in\mathcal{A}}|X_{ia}-X_{ja}|^q\right)^{1/q}\right]\right)^2}{\mathrm{E}[(X_a^{\max}-X_a^{\min})^2]} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}}+\mu^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^{\max}-X_a^{\min})^2]} - \frac{\mu^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^{\max}-X_a^{\min})^2]} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^{\max}-X_a^{\min})^2]} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}}}{\mathrm{E}[(X_a^{\max})^2]-2\mathrm{E}(X_a^{\max})\mathrm{E}(X_a^{\min})+\mathrm{E}(X_a^{\min})} \\
&= \frac{\sigma^2_{D_{ij}^{(q)}}}{\mu^{(2)}_{\max}(m)-2\mu^{(1)}_{\max}(m)\mu^{(1)}_{\min}(m)+\mu^{(2)}_{\min}(m)},
\end{aligned}
\tag{73}
$$

where m is the size of the sample from which extremes are derived, $\mu^{(1)}_{\max}(m)$ is the expected value of the sample maximum (Eq. 62), and $\mu^{(1)}_{\min}$ is the expected value of the sample minimum.

With the mean (Eq. 72) and variance (Eq. 73) of the max-min normalized distance (Eq. 55), we have the following generalized estimate for the asymptotic distribution of the max-min normalized distance distribution

$$
D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{\mu^{(1)}_{\max}(m)-\mu^{(1)}_{\min}(m)}, \frac{\sigma^2_{D_{ij}^{(q)}}}{\mu^{(2)}_{\max}(m)-2\mu^{(1)}_{\max}(m)\mu^{(1)}_{\min}(m)+\mu^{(2)}_{\min}(m)}\right), \tag{74}
$$

where m is the size of the sample from which extremes are derived, $\mu^{(1)}_{\max}(m)$ is the expected value of the sample maximum (Eq. 62), and $\mu^{(1)}_{\min}$ is the expected value of the sample minimum.

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 expected max-min normalized pairwise distance (Eq. 72) exclusively in terms of the expected sample maximum. This result is given by the following

$$
\mu_{D_{ij}^{(q*)}} \approx \frac{\mu_{D_{ij}^{(q)}}}{2\mu^{(1)}_{\max}(m)}, \tag{75}
$$

where m is the size of the sample from which the maximum is derived and $\mu^{(1)}_{\max}(m)$ is the expected value of the sample maximum (Eq. 62).

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

normalized distance distribution

$$\begin{aligned}\sigma^2_{D_{ij}^{(q*)}} &\approx \frac{\sigma^2_{D_{ij}^{(q)}}}{2\mu^{(2)}_{\max}(m) + 2\left[\mu^{(1)}_{\max}(m)\right]^2} \\ &= \frac{\sigma^2_{D_{ij}^{(q)}}}{2\left(\sigma^2_{\max}(m) + \left[\mu^{(1)}_{\max}(m)\right]^2\right)},\end{aligned} \tag{76}$$

where m is the size of the sample from which extremes are derived, $\mu^{(1)}_{\max}(m)$ is the expected value of the sample maximum (Eq. 62), and $\sigma^2_{\max}(m)$ is the variance of the sample maximum (Eq. 64).

Therefore, the asymptotic distribution of the max-min normalized distance distribution (Eq. 74) becomes

$$D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu^{(1)}_{\max}(m)}, \frac{\sigma^2_{D_{ij}^{(q)}}}{2\left(\sigma^2_{\max}(m) + \left[\mu^{(1)}_{\max}(m)\right]^2\right)}\right), \tag{77}$$

where m is the size of the sample from which extremes are derived, $\mu^{(1)}_{\max}(m)$ is the expected value of the sample maximum (Eq. 62), and $\sigma^2_{\max}(m)$ is the variance of the sample maximum (Eq. 64).

We have now derived asymptotic estimates of the moments of the max-min normalized L_q distance metric (Eq. 55) for any continuous data distribution. In the next two sections, we examine the max-min normalized L_q distance on standard normal and standard uniform data. As in previous sections in which we analyzed the standard L_q metric (Eq. 1), we will use the more general results for the max-min L_q metric to derive asymptotic estimates for normalized Manhattan ($q = 1$) and Euclidean ($q = 2$).

2.4.1 Standard normal data

The standard normal distribution has zero skewness, infinite support that is symmetric about 0, and all moments. This implies that the corresponding 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 (Eqs. 62 and 63) 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 $b_m = \frac{1}{c_m}$, where Φ is the standard normal cumulative distribution function. Using Taylor's Theorem, we have the following expansion

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

where m is the size of the sample from which the maximum is derived.

In order to simplify the right-hand side of this expansion (Eq. 78), we will use the well known Mills Ratio Bounds [17] given by the following

$$1 \leq \frac{\phi(x)}{x\Phi(-x)} \leq 1 + \frac{1}{x^2} \quad , x > 0, \tag{79}$$

where Φ and ϕ once again represent the cumulative distribution function and density function, respecively, of the standard normal distribution.

The inequalities given above (Eq. 79) show that

$$\frac{\phi(x)}{x\Phi(-x)} \to 1 \text{ as } x \to \infty.$$

This further implies that

$$\frac{\phi(c_m)}{c_m\Phi(-c_m)} \to 1 \text{ as } m \to \infty$$

since

$$c_m = -\Phi^{-1}\left(\frac{1}{m}\right) \to \infty \text{ as } m \to \infty.$$

This gives us the following approximation of the right-hand side of the expansion (Eq. 78) given previously

$$\begin{aligned} &\log\Phi(-c_m - b_m x) \approx \log\left(\frac{1}{m}\right) - x + \mathcal{O}(b_m^2 x^2) \\ &\Rightarrow \Phi(-c_m - b_m x) \approx \frac{1}{m} e^{-x+\mathcal{O}(b_m^2 x^2)} \\ &\Rightarrow \Phi(c_m + b_m x) \approx 1 - \frac{1}{m} e^{-x+\mathcal{O}(b_m^2 x^2)}, \end{aligned} \tag{80}$$

where m is the size of the sample from which the maximum is derived.

Using the approximation of expansion given previously (Eq. 80), we now derive the limit distribution for the sample maximum in standard normal data as

$$\begin{aligned} \mathrm{P}\left(\frac{X_a^{\max} - c_m}{b_m} \le x\right) &= \mathrm{P}(X_a^{\max} \le c_m + b_m x) \\ &= \Phi^m(c_m + b_m x) \\ &\approx \left(1 - \frac{1}{m} e^{-x+\mathcal{O}(b_m^2 x^2)}\right)^m \\ &= \left(1 - \frac{1}{m} e^{-x+\mathcal{O}\left(\frac{1}{c_m^2} x^2\right)}\right)^m \\ &\approx \left(1 - \frac{1}{m} e^{-x}\right)^m \\ \Rightarrow \lim_{m\to\infty} \mathrm{P}\left(\frac{X_a^{\max} - c_m}{b_m} \le x\right) &= \lim_{m\to\infty} \left(1 - \frac{1}{m} e^{-x}\right)^m \\ &= e^{-e^{-x}}, \end{aligned} \tag{81}$$

which is the cumulative distribution function of the standard Gumbel distribution. The mean of this distribution is given by the following

$$\mathrm{E}(X_a^{\max}) = \mu_{\max}^{(1)} = -\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)}, \tag{82}$$

where m is the size of the sample from which the maximum is derived and γ is the well known Euler-Mascheroni constant. This constant has many equivalent definitions, one of which is given by

$$\gamma = \lim_{m\to\infty} \left(-\log(m) + \sum_{k=1}^{m} \frac{1}{k}\right).$$

Perhaps a more convenient definition of the Euler-Mascheroni constant is simply

$$\gamma = -\Gamma'(1) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\int_0^\infty z^{t-1}e^{-z}\mathrm{d}z\right)\Big|_{t=1},$$

which is just the additive inverse of the first derivative of the gamma function evaluated at 1.

The median of the distribution of the maximum for standard normal data is given by

$$\widetilde{\mu}_{\max} = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right), \tag{83}$$

where m is the size of the sample from which the maximum is derived.

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

$$\mathrm{Var}(X_a^{\max}) = \frac{\pi^2}{6}\left(\frac{1}{-\Phi^{-1}\left(\frac{1}{m}\right)}\right)^2, \tag{84}$$

where m is the size of the sample from which the maximum is derived.

For typical sample sizes m in high-dimensional spaces, the variance estimate (Eq. 84) exceeds the variance of the sample maximum significantly. Using the fact that

$$-\Phi^{-1}\left(\frac{1}{m}\right) \dot{\sim} \sqrt{2\log(m)} \text{ [18]}$$

and

$$\frac{1}{2\log(m)} \leq \left(\frac{1}{-\Phi^{-1}\left(\frac{1}{m}\right)}\right)^2, \quad m \geq 2,$$

we can get a more accurate approximation of the variance with the following

$$\begin{aligned}\sigma^2_{\max}(m) = \mathrm{Var}(X_a^{\max}) &\approx \frac{\pi^2}{6}\left(\frac{1}{\sqrt{2\log(m)}}\right)^2 \\ &= \frac{\pi^2}{12\log(m)}.\end{aligned} \tag{85}$$

Therefore, the mean of the range of m iid standard normal random variables is given by

$$\mathrm{E}(X_a^{\max} - X_a^{\min}) = 2\mu^{(1)}_{\max}(m) = 2\left[-\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)}\right], \tag{86}$$

where γ is the Euler-Mascheroni constant.

It is well known that the sample extremes from the standard normal distribution are approximately uncorrelated for large sample size m [16]. This implies that we can approximate the variance of the range of m iid standard normal random variables with the following result

$$\begin{aligned}\mathrm{Var}(X_a^{\max} - X_a^{\min}) &\approx \mathrm{Var}(X_a^{\max}) + \mathrm{Var}(X_a^{\min}) \\ &= \sigma^2_{\max}(m) + \sigma^2_{\min}(m) \\ &= 2\sigma^2_{\max}(m) \\ &\approx 2\left(\frac{\pi^2}{12\log(m)}\right) \\ &= \frac{\pi^2}{6\log(m)}.\end{aligned} \tag{87}$$

For the purpose of approximating the mean and variance of the max-min normalized distance distribution, we observe empirically that the formula for the median of the distribution of the attribute maximum (Eq. 83) yields more accurate results. More precisely, the approximation of the expected maximum (Eq. 82) overestimates the sample maximum slightly. The formula for the median of the sample maximum (Eq. 83) provides a more accurate estimate of this sample extreme. Therefore, the following estimate for the mean of the attribute range will be used instead

$$\mathrm{E}(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m) \approx 2\left[\frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)\right], \tag{88}$$

where m is the size of the sample from which extremes are derived.

We have already determined the mean and variance (Eq. 27) for the L_q metric (Eq. 1) on standard normal data. Using the expected value of the sample maximum (Eq. 88), the variance of the sample maximum (Eq. 87), and the general formulas for the mean and variance of the max-min normalized distance distribution (Eq. 77), this leads us to the following asymptotic estimate for the distribution of the max-min normalized distances for standard normal data

$$D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}, \frac{6\log(m)\sigma^2_{D_{ij}^{(q)}}}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m)\right]^2\log(m)}\right). \tag{89}$$

where m is the size of the sample from which the maximum is derived, $\mu_{\max}^{(1)}$ is the median of the sample maximum (Eq. 83), $\mu_{D_{ij}^{(q)}}$ is the expected L_q pairwise distance (Eq. 25), and $\sigma^2_{D_{ij}^{(q)}}$ is the variance of the L_q pairwise distance (Eq. 26). The moments of the max-min normalized L_q distance metric in standard normal data (Eq. 89) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 1).

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 (Eq. 57) and minimum (Eq. 60), we have the following distribution functions for standard uniform data

$$F_{\max}(x) = x^m \tag{90}$$

and

$$F_{\min}(x) = 1 - (1-x)^m, \tag{91}$$

where m is the size of the sample from which extremes are derived.

Using the general forms of the density functions of the maximum (Eq. 58) and minimum (Eq. 61), we have the following density functions for standard uniform data

$$f_{\max}(x) = mx^{m-1} \tag{92}$$

and

$$f_{\min}(x) = m(1-x)^{m-1}, \tag{93}$$

where m is the size of the sample from which extremes are derived.

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

$$\begin{aligned} \mathrm{E}(X_a^{\max}) = \mu_{\max}^{(1)}(m) &= \int_0^1 x f_{\max}(x)\mathrm{d}x \\ &= \int_0^1 x[mx^{m-1}]\mathrm{d}x \\ &= \frac{m}{m+1} \end{aligned} \tag{94}$$

and

$$\begin{aligned} \mathrm{E}(X_a^{\min}) = \mu_{\min}^{(1)}(m) &= \int_0^1 x f_{\min}(x)\mathrm{d}x \\ &= \int_0^1 x[m(1-x)^{m-1}]\mathrm{d}x \\ &= \frac{1}{m+1}, \end{aligned} \tag{95}$$

where m is the size of the sample from which extremes are derived.

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

$$\begin{aligned} \mathrm{E}[(X_a^{\max} - X_a^{\min})^2] &= \mathrm{E}[(X_a^{\max})^2 - 2X_a^{\max}X_a^{\min} + (X_a^{\min})^2] \\ &= \mathrm{E}[(X_a^{\max})^2] - 2\mathrm{E}(X_a^{\max})\mathrm{E}(X_a^{\min}) + \mathrm{E}[(X_a^{\min})^2] \\ &= \mu_{\max}^{(2)}(m) - 2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m) + \mu_{\min}^{(2)}(m) \\ &= \int_0^1 x^2[mx^{m-1}]\mathrm{d}x - 2\left(\frac{m}{m+1}\right)\left(\frac{1}{m+1}\right) \\ &\qquad + \int_0^1 x^2[m(1-x)^{m-1}]\mathrm{d}x \\ &= \frac{m}{m+2} - \frac{2m}{(m+1)^2} + \frac{2}{(m+1)(m+2)} \\ &= \frac{m^3 - m + 2}{(m+2)(m+1)^2}, \end{aligned} \tag{96}$$

where m is the size of the sample from which extremes are derived.

Using the general asymptotic distribution of max-min normalized distances for any data type (Eq. 74) and the mean and variance (Eq. 37) of the standard L_q distance metric (Eq. 1), we have the following asymptotic estimate for the max-min normalized distance distribution for standard uniform data

$$D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{(m+1)\mu_{D_{ij}^{(q)}}}{m-1}, \frac{(m+2)(m+1)^2\sigma^2_{D_{ij}^{(q)}}}{m^3-m+2}\right), \tag{97}$$

where m is the size of the sample from which extremes are derived, $\mu_{D_{ij}^{(q)}}$ is the expected value (Eq. 35) of the L_q metric (Eq. 1) in standard uniform data, and $\sigma^2_{D_{ij}^{(q)}}$ is the variance (Eq. 36) of the L_q metric (Eq. 1) in standard uniform data. The moments of the max-min normalized L_q distance metric in standard uniform data (Eq. 89) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 1).

2.5 Normalized Manhattan ($q = 1$)

Using the general asymptotic results for mean and variance of max-min normalized distances in standard normal and standard uniform data (Eqs. 89 and 97) for any value of $q \in \mathbb{N}$, we can substitute a particular value of q in order to determine a more specified distribution for the normalized distance (Eq. 55). 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 the asymptotic formula for the expected max-min normalized distance (Eq. 89), we derive the expected normalized Manhattan distance in standard normal data as follows

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(1*)}\right) &= \frac{\mu_{D_{ij}^{(1)}}}{2\mu_{\max}^{(1)}(m)} \\ &= \frac{p}{\sqrt{\pi}\mu_{\max}^{(1)}(m)}, \end{aligned} \tag{98}$$

where $\mu_{\max}^{(1)}(m)$ is the expected attribute maximum (Eq. 83),m is the size of the sample from which the maximum is derived, and p is the total number of attributes.

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

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(1*)}\right) &= \frac{6\log(m)\sigma^2_{D_{ij}^{(1)}}}{\pi^2 + 24\left[\mu_{\max}^{(1)}\right]^2\log(m)} \\ &= \frac{12p(\pi-2)\log(m)}{\pi\left(\pi^2 + 24\left[\mu_{\max}^{(1)}\right]^2\log(m)\right)}, \end{aligned} \tag{99}$$

where $\mu_{\max}^{(1)}(m)$ is the expected attribute maximum (Eq. 83), m is the size of the sample from which the maximum is derived, and p is the total number of attributes. Similar to the variance of the standard Manhattan distance, the variance of the max-min normalized Manhattan distance is on the order of p for fixed instance dimension m. For fixed p, the variance (Eq. 99) vanishes as m grows without bound. If we fix m, the same variance increases monotonically with increasing p. The moments derived in this section (Eqs. 98 and 99) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 3).

2.5.2 Standard uniform data

Substituting $q = 1$ into the asymptotic formula for the expected max-min pairwise distance (Eq. 97), we derive the expected normalized Manhattan distance in standard uniform data as

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(1*)}\right) &= \frac{(m+1)\mu_{D_{ij}^{(1)}}}{m-1} \\ &= \frac{(m+1)p}{3(m-1)}, \end{aligned} \tag{100}$$

where m is the size of the sample from which extremes are derived and p is the total number attributes.

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

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(1*)}\right) &= \frac{(m+2)(m+1)^2\sigma^2_{D_{ij}^{(1)}}}{m^3-m+2} \\ &= \frac{(m+2)(m+1)^2 p}{18(m^3-m+2)}, \end{aligned} \tag{101}$$

where m is the size of the sample from which extremes are derived and p is the total number of attributes. Interestingly, the variance of the max-min normalized Manhattan distance in standard uniform data approaches $p/18$ as m increases without bound for a fixed number of features p. This is the same asymptotic value to which the variance of the standard Manhattan distance (Eq. 40) converges. Therefore, large sample sizes make the variance of the normalized Manhattan distance approach the variance of the standard Manhattan distance in standard uniform data. The moments derived in this section (Eqs. 100 and 101) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 3).

2.6 Normalized Euclidean ($q = 2$)

Analogous to the previous section, we demonstrate the use the asymptotic moment estimates for max-min normalized distances (Eq. 89 and 97) for the max-min normalized Euclidean ($q = 2$) metric (Eq. 55) for both standard normal and standard uniform data.

2.6.1 Standard normal data

Substituting $q = 2$ into the asymptotic formula for the expected max-min normalized pairwise distance (Eq. 89), we derive the expected normalized Euclidean distance in standard normal data as

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2*)}\right) &= \frac{\mu^{D_{ij}^{(2)}}}{2\mu^{(1)}_{\max}(m)} \\ &= \frac{\sqrt{2p-1}}{2\mu^{(1)}_{\max}(m)}, \end{aligned} \tag{102}$$

where $\mu^{(1)}_{\max}(m)$ is the expected attribute maximum (Eq. 83), m is the size of the sample from which the maximum is derived, and p is the total number of attributes.

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

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(2*)}\right) &= \frac{6\log(m)\sigma^2_{D_{ij}^{(2)}}}{\pi^2 + 24\left[\mu^{(1)}_{\max}(m)\right]^2\log(m)} \\ &= \frac{6\log(m)}{\pi^2 + 24\left[\mu^{(1)}_{\max}(m)\right]^2\log(m)}, \end{aligned} \tag{103}$$

where $\mu^{(1)}_{\max}(m)$ is the expected attribute maximum (Eq. 83) and m is the size of the sample from which the maximum is derived. It is interesting to note that the variance (Eq. 103) vanishes as the sample size m increases without bound, which means that all distances will be tightly clustered about the mean (Eq. 102). This is different than the variance of the standard L_2 metric (Eq. 49), which is asymptotically equal to 1. This could imply that any two pairwise distances computed with the max-min normalized Euclidean metric in a large sample space m may be indistinguishable, which is another curse of dimensionality. The moments derived in this section (Eqs. 102 and 103) are

summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 3).

2.6.2 Standard uniform data

Substituting $q = 2$ into the asymptotic formula for the expected max-min normalized pairwise distance (Eq. 97), we derive the expected normalized Euclidean distance in standard uniform data as

$$\begin{aligned} \mathrm{E}\left(D_{ij}^{(2*)}\right) &= \frac{(m+1)\mu_{D_{ij}^{(2)}}}{m-1} \\ &= \sqrt{\frac{p}{6} - \frac{7}{120}}\left(\frac{m+1}{m-1}\right). \end{aligned} \tag{104}$$

where m is the size of the sample from which extremes are derived and p is the total number of attributes.

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

$$\begin{aligned} \mathrm{Var}\left(D_{ij}^{(2*)}\right) &= \frac{(m+2)(m+1)^2\sigma^2_{D_{ij}^{(2)}}}{m^3-m+2} \\ &= \frac{7(m+2)(m+1)^2}{120(m^3-m+2)}. \end{aligned} \tag{105}$$

where m is the size of the sample from which extremes are derived. Similar to the variance of max-min normalized Manhattan distances in standard uniform data (Eq. 101), the variance of normalized Euclidean distances approaches the variance of the standard Euclidean distances in uniform data (Eq. 52) as m increases without bound. That is, the variance of the max-min normalized Euclidean distance (Eq. 105) approaches $7/120$ as m grows larger. The moments derived in this section (Eqs. 104 and 105) are summarized in a table that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Tab. 3).

We conclude this section with summary tables (Tabs. 1-3) that contain all of our asymptotic results for both standard and max-min normalized L_q metrics in each data type we have considered. This includes our most general results for any combination of sample size m, number of features p, type of metric L_q, and data type (Tab. 1). From these more general derivations, we show the results of the standard L_1 and L_2 metrics for any combination of sample size m, number of features p, and data type (Tab. 2). Our last set of summarized results show asymptotics for the max-min normalized L_1 and L_2 metrics for any combination of sample size m, number of features p, and data type (Tab. 3). For both standard and max-min normalized L_2 metrics (Tabs. 2 and 3), the low-dimensional improved estimates of sample means (Eqs. 50 and 53) are used because they perform well at both low and high attribute dimension p.

In the next section, we make a transition into discrete GWAS data. We will discuss some commonly known metrics and then a relatively new metric, which will lead us into novel asymptotic results for this data type.

Table 1. Summary of distance distribution derivations for standard normal and standard uniform data. Asymptotic estimates are given for both standard (Eq. 1) and max-min normalized (Eq. 55) q-metrics. These estimates are relevant for all $q \in \mathbb{N}$ and p for which the normality assumption of distances holds.

***q*-Metric**	**Data**	**Stat**	**Formula (Eq. #)**
standard (**Eq. 1**)	$\mathcal{N}(0,1)$	mean	$\left(\frac{2^q\Gamma\left(\frac{q+1}{2}\right)p}{\sqrt{\pi}}\right)^{1/q}$ (**25**)
	$\mathcal{N}(0,1)$	variance	$\frac{4^q p}{q^2\left(\frac{2^q\Gamma\left(\frac{1}{2}q+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}}-\frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]$ (**26**)
	$\mathcal{U}(0,1)$	mean	$\left(\frac{2p}{(q+2)(q+1)}\right)^{1/q}$ (**35**)
	$\mathcal{U}(0,1)$	variance	$\frac{p}{q^2\left(\frac{2p}{(q+2)(q+1)}\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{1}{(q+1)(2q+1)}-\left(\frac{2}{(q+2)(q+1)}\right)^2\right]$ (**36**)
max-min normalized (**Eq. 55**)	$\mathcal{N}(0,1)$	mean	$\boxed{\frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}}$ (**89**) where $\mu_{D_{ij}^{(q)}}$ and $\mu_{\max}^{(1)}(m)$ are given by Eqs. 25 and 83 respectively.
	$\mathcal{N}(0,1)$	variance	$\boxed{\frac{6\log(m)\sigma^2_{D_{ij}^{(q)}}}{\pi^2+24\left[\mu_{\max}^{(1)}(m)\right]^2\log(m)}}$ (**89**) where $\sigma^2_{D_{ij}^{(q)}}$ and $\mu_{\max}^{(1)}(m)$ are given by Eqs. 25 and 83 respectively.
	$\mathcal{U}(0,1)$	mean	$\boxed{\frac{(m+1)\mu_{D_{ij}^{(q)}}}{m-1}}$ (**97**) where $\mu_{D_{ij}^{(q)}}$ is given by Eq. 35
	$\mathcal{U}(0,1)$	variance	$\boxed{\frac{(m+2)(m+1)^2\sigma^2_{D_{ij}^{(q)}}}{m^3-m+2}}$ (**97**) where $\sigma^2_{D_{ij}^{(q)}}$ is given by Eq. 36

Table 2. Asymptotic estimates of 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 $\boldsymbol{L_1}$ **(Eq. 1)**	$\mathcal{N}(0,1)$	mean	$\frac{2p}{\sqrt{\pi}}$ (**38**)
		variance	$\frac{2(\pi-2)p}{\pi}$ (**39**)
	$\mathcal{U}(0,1)$	mean	$\frac{p}{3}$ (**40**)
		variance	$\frac{p}{18}$ (**41**)
standard $\boldsymbol{L_2}$ **(Eq. 1)**	$\mathcal{N}(0,1)$	mean	$\sqrt{2p-1}$ (**48**)
		variance	1 (**49**)
	$\mathcal{U}(0,1)$	mean	$\sqrt{\frac{p}{6}-\frac{7}{120}}$ (**51**)
		variance	$\frac{7}{120}$ (**52**)

Table 3. Asymptotic estimates of 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. The cumulative distribution function of the standard normal distribution is represented by Φ. Furthermore, $\mu^{(1)}_{\max}(m)$ (Eq. 83) is the asymptotic median of the sample maximum from m standard normal random samples.

***q*-Metric**	**Data**	**Stat**	**Formula (Eq. #)**
max-min normalized $\boldsymbol{L_1}$ (**Eq. 55**)	$\mathcal{N}(0,1)$	mean	$\boxed{\frac{p}{\sqrt{\pi}\mu^{(1)}_{\max}(m)}}$ (**98**) where $\mu^{(1)}_{\max}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
		variance	$\boxed{\frac{12p(\pi-2)\log(m)}{\pi\left(\pi^2+24\left[\mu^{(1)}_{\max}(m)\right]^2\log(m)\right)}}$ (**99**) where $\mu^{(1)}_{\max}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
	$\mathcal{U}(0,1)$	mean	$\frac{(m+1)p}{3(m-1)}$ (**100**)
		variance	$\frac{(m+2)(m+1)^2p}{18(m^3-m+2)}$ (**100**)
max-min normalized $\boldsymbol{L_2}$ (**Eq. 55**)	$\mathcal{N}(0,1)$	mean	$\boxed{\frac{\sqrt{2p-1}}{2\mu^{(1)}_{\max}(m)}}$ (**102**) where $\mu^{(1)}_{\max}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
		variance	$\boxed{\frac{6\log(m)}{\pi^2+24\left[\mu^{(1)}_{\max}(m)\right]^2\log(m)}}$ (**103**) where $\mu^{(1)}_{\max}(m) = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)$
	$\mathcal{U}(0,1)$	mean	$\sqrt{\frac{p}{6} - \frac{7}{120}}\left(\frac{m+1}{m-1}\right)$ (**104**)
		variance	$\frac{7(m+2)(m+1)^2}{120(m^3-m+2)}$ (**105**)

3 GWAS distance distributions

In genome-wide association studies, data is encoded by the minor allele at a particular locus a, which is just the second most common allele (adenine-A, thymine-T, cytosine-C,

or guanine-G) associated with a given feature a in the data set. Features in GWAS data are single nucleotide polymorphisms (SNPs), or mutations involving the substitution, deletion, or insertion of one nucleotide at some point in the DNA sequence of an organism. These are common mutations that can affect how an individual reacts to certain pathogens or the susceptibility for certain diseases. Feature selection in GWAS is typically concerned with finding interacting SNPs that are associated with disease susceptibility [19].

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

$$X_{ia} = \begin{cases} 0 & \text{if there are no minor alleles at locus } a, \\ 1 & \text{if there is 1 minor allele at locus } a, \\ 2 & \text{if there are 2 minor alleles at locus } a. \end{cases} \tag{106}$$

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 distance metrics for GWAS data are the Genotype Mismatch (GM) and Allele Mismatch (AM) metrics. The GM and AM metrics are defined by

$$\mathrm{d}_{ij}^{\mathrm{GM}}(a) = \begin{cases} 0 & \text{if } X_{ia} \neq X_{ja}, \\ 1 & \text{otherwise} \end{cases} \tag{107}$$

and

$$\mathrm{d}_{ij}^{\mathrm{AM}}(a) = \frac{1}{2}\big|X_{ia} - X_{ja}\big|. \tag{108}$$

A more informative metric must take into account whether differences in allele frequency at a particular locus a result from transitions or transversions. A new discrete metric that accounts for transitions (Ti) and transversions (Tv) was introduced in [4]. This metric is given by the following

$$\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = \begin{cases} 0 & \text{if } X_{ia} = X_{ja} \text{ and Ti/Tv}, \\ 1/4 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Ti}, \\ 1/2 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Tv}, \\ 3/4 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Ti}, \\ 1 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Tv}. \end{cases} \tag{109}$$

With these GWAS distance metrics, we then compute the pairwise distance between two instances $i, j \in \mathcal{I}$ with

$$D_{ij}^{\mathrm{GM}}(a) = \sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\mathrm{GM}}(a), \tag{110}$$

$$D_{ij}^{\mathrm{AM}}(a) = \sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\mathrm{AM}}(a), \text{ or} \tag{111}$$

$$D_{ij}^{\mathrm{TiTv}}(a) = \sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\mathrm{TiTv}}(a). \tag{112}$$

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 (Eqs. 107-109) are asymptotically normal. With this Gaussian limiting behavior, we will proceed by deriving the mean and variance for each distance distribution induced by these three GWAS metrics.

3.1 GM distance distribution

The simplest distance metric in nearest-neighbor feature selection in GWAS data is the GM distance metric (Eq. 110). The GM attribute diff (Eq. 107) indicates only whether two genotypes are the same or not. There are many ways two genotypes could differ, but this metric does not record this information. We will now derive the moments for the GM distance (Eq. 110), which are sufficient for defining its corresponding asymptotic distribution.

The expected value of the GM attribute diff metric (Eq. 107) is given by the following

$$\begin{aligned}
\mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right] &= \sum_{k=0}^{1} k \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = k\right] \\
&= 0 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 0\right] + 1 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 1\right] \\
&= \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 1\right] \\
&= 2\mathrm{P}[X_{ia} = 0, X_{ja} = 1] + 2\mathrm{P}[X_{ia} = 1, X_{ja} = 2] + 2\mathrm{P}[X_{ia} = 0, X_{ja} = 2] \\
&= 4(1-f_a)^3 f_a + 4(1-f_a) f_a^3 + 2(1-f_a)^2 f_a^2 \\
&= 2\left[2(1-f_a)^3 f_a + 2(1-f_a) f_a^3 + (1-f_a)^2 f_a^2\right] \\
&= 2F^{\mathrm{GM}}(a),
\end{aligned} \tag{113}$$

where $F^{\mathrm{GM}}(a) = 2(1-f_a)^3 f_a + 2(1-f_a) f_a^3 + (1-f_a)^2 f_a^2$ and f_a is the probability of a minor allele occurring at locus a.

Then the expected pairwise GM distance between instances $i, j \in \mathcal{I}$ is given by

$$\begin{aligned}
\mathrm{E}\left(D_{ij}^{\mathrm{GM}}\right) &= \mathrm{E}\left(\sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\mathrm{GM}}(a)\right) \\
&= \sum_{a \in \mathcal{A}} \mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right] \\
&= 2 \sum_{a \in \mathcal{A}} F^{\mathrm{GM}}(a),
\end{aligned} \tag{114}$$

where $F^{\mathrm{GM}}(a) = 2(1-f_a)^3 f_a + 2(1-f_a) f_a^3 + (1-f_a)^2 f_a^2$ and f_a is the probability of a minor allele occurring at locus a. We see that the expected GM pairwise distance (Eq. 114) relies only on the minor allele probabilities f_a for all $a \in \mathcal{A}$. In real data, we can easily determine these probabilities by dividing the total number of minor alleles at locus a by the twice the number of instances m. To be more explicit, this is just

$$f_a = \frac{1}{2m} \sum_{i \in \mathcal{I}} X_{ia}, \quad \text{for all } a \in \mathcal{A},$$

where m is the number of instances (or sample size). This is because each instance has two alleles, the minor major alleles, at each locus. Therefore, the total number of alleles at locus a is $2m$.

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

$$\begin{aligned}
\mathrm{E}\left[\left(D_{ij}^{\mathrm{GM}}\right)^2\right] &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\mathrm{GM}}(a)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a\in\mathcal{A}}\left(\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r\in\mathcal{A}}\sum_{s\leq r-1} \mathrm{d}_{ij}^{\mathrm{GM}}(r)\cdot \mathrm{d}_{ij}^{\mathrm{GM}}(s)\right] \\
&= \sum_{a\in\mathcal{A}}\left(\sum_{k=0}^{1} k^2 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a)=k\right]\right) \\
&+ 2\sum_{a\in\mathcal{A}}\sum_{s\leq r-1}\left(\sum_{k=0}^{1} k\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(r)=k\right]\right)\cdot\left(\sum_{k=0}^{1} k\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(s)=k\right]\right) \\
&= 2\sum_{a\in\mathcal{A}} F^{\mathrm{GM}}(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F^{\mathrm{GM}}(\lambda),
\end{aligned} \tag{115}$$

where $F^{\mathrm{GM}}(a) = 2(1-f_a)^3 f_a + 2(1-f_a)f_a^3 + (1-f_a)^2 f_a^2$ and f_a is the probability of a minor allele occurring at locus a.

Using the first (Eq. 114) and second (Eq. 115) raw moments of the GM distance, the variance is given by

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

where $F^{\mathrm{GM}}(a) = 2(1-f_a)^3 f_a + 2(1-f_a)f_a^3 + (1-f_a)^2 f_a^2$ and f_a is the probability of a minor allele occurring at locus a. Hence, the variance of the asymptotic GM distance distribution also just depends on the minor allele probabilities f_a for all $a \in \mathcal{A}$. This implies that the limiting GM distance distribution is fully determined by the minor allele probabilites, which are known in real data.

With the mean and variance estimates (Eqs. 114 and 116), the asymptotic GM distance distribution is given by the following

$$D_{ij}^{\mathrm{GM}} \dot{\sim} \mathcal{N}\left(2\sum_{a\in\mathcal{A}} F^{\mathrm{GM}}(a), 2\sum_{a\in\mathcal{A}} F^{\mathrm{GM}}(a)[1-2F^{\mathrm{GM}}(a)]\right), \tag{117}$$

where $F^{\mathrm{GM}}(a) = 2(1-f_a)^3 f_a + 2(1-f_a)f_a^3 + (1-f_a)^2 f_a^2$ and f_a is the probability of a minor allele occurring at locus a.

We have concluded our treatment of GM distances in random GWAS data, where the independence assumption of holds for minor allele probabilites f_a and binomial samples $X_{ia} \sim \mathcal{B}(2, f_a)$ for all $a \in \mathcal{A}$. In the next section, we consider the AM distance distribution in GWAS data. This metric incorporates genotype differences at the level of individual allelic differences, so it is more informative than the GM metric.

3.2 AM distance distribution

As we have mentioned previously, the AM attribute diff metric (Eq. 108) is slightly more dynamic than the GM metric because the AM metric accounts for differences between the alleles of two genotypes. In this section, we derive moments of the AM distance metric (Eq. 111) that adequately define its corresponding asymptotic distribution.

The expected value of the AM attribute diff metric (Eq. 108) is given by the following

$$\begin{aligned}
\mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a)\right] &= \sum_{k\in\mathcal{D}} k \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a) = k\right] \\
&= 0 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a) = 0\right] + \frac{1}{2} \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a) = \frac{1}{2}\right] + 1 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a) = 1\right] \\
&= \frac{1}{2}\left(2\mathrm{P}\left[X_{ia} = 0, X_{ja} = 1\right] + 2\mathrm{P}\left[X_{ia} = 1, X_{ja} = 2\right]\right) \\
&+ 2\mathrm{P}\left[X_{ia} = 0, X_{ja} = 2\right] \\
&= \mathrm{P}\left[X_{ia} = 0, X_{ja} = 1\right] + \mathrm{P}\left[X_{ia} = 1, X_{ja} = 2\right] + 2\mathrm{P}\left[X_{ia} = 0, X_{ja} = 2\right] \\
&= 2(1-f_a)^3 f_a + 2(1-f_a) f_a^3 + 2(1-f_a)^2 f_a^2 \\
&= 2\left[(1-f_a)^3 f_a + (1-f_a) f_a^3 + (1-f_a)^2 f_a^2\right] \\
&= 2F^{\mathrm{AM}}(a),
\end{aligned} \tag{118}$$

where $F^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + (1-f_a) f_a^3 + (1-f_a)^2 f_a^2$, $\mathcal{D} = \{0, 1/2, 1\}$, and f_a is the probability of a minor allele occurring at locus a.

Using the expected AM attribute diff (Eq. 118), the expected pairwise AM distance (Eq. 111) between instances $i, j \in \mathcal{I}$ is given by

$$\begin{aligned}
\mathrm{E}\left(D_{ij}^{\mathrm{AM}}\right) &= \mathrm{E}\left(\sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\mathrm{AM}}(a)\right) \\
&= \sum_{a\in\mathcal{A}} \mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a)\right] \\
&= 2\sum_{a\in\mathcal{A}} F^{\mathrm{AM}}(a).
\end{aligned} \tag{119}$$

where $F^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + (1-f_a) f_a^3 + (1-f_a)^2 f_a^2$ and f_a is the probability of a minor allele occurring at locus a. Similar to GM distances, the expected AM distance (Eq. 119) depends only on the minor allele probabilities f_a for all $a \in \mathcal{A}$. This is to be expected because, although the AM metric is more informative, it still only accounts for simple differences between nucleotides of two instances $i, j \in \mathcal{I}$ at some locus a.

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

$$\begin{aligned}
\mathrm{E}\left[\left(D_{ij}^{\mathrm{AM}}\right)^2\right] &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\mathrm{AM}}(a)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a\in\mathcal{A}}\left(\mathrm{d}_{ij}^{\mathrm{AM}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r\in\mathcal{A}}\sum_{s\leq r-1} \mathrm{d}_{ij}^{\mathrm{AM}}(r)\cdot \mathrm{d}_{ij}^{\mathrm{AM}}(s)\right] \\
&= \sum_{a\in\mathcal{A}}\left(\sum_{k\in\mathcal{D}} k^2\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a)=k\right]\right) \\
&+ 2\sum_{a\in\mathcal{A}}\sum_{s\leq r-1}\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(r)=k\right]\right)\cdot\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(s)=k\right]\right) \\
&= \sum_{a\in\mathcal{A}} G^{\mathrm{AM}}(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F^{\mathrm{AM}}(\lambda),
\end{aligned} \tag{120}$$

where $G^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$, $F^{\mathrm{AM}}(\lambda) = (1-f_\lambda)^3 f_\lambda + f_\lambda^3(1-f_\lambda) + (1-f_\lambda)^2 f_\lambda^2$, and f_a is the probability of a minor allele occurring at locus a.

Using the first (Eq. 119) and second (Eq. 120) raw moments of the asymptotic AM distance distribution, the variance is given by

$$\begin{aligned}
\mathrm{Var}\left(D_{ij}^{\mathrm{AM}}\right) &= \mathrm{E}\left[\left(D_{ij}^{\mathrm{AM}}\right)^2\right] - \left[\mathrm{E}\left(D_{ij}^{\mathrm{AM}}\right)\right]^2 \\
&= \sum_{a\in\mathcal{A}} G^{\mathrm{AM}}(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F^{\mathrm{AM}}(\lambda) - 4\left(\sum_{a\in\mathcal{A}} F^{\mathrm{AM}}(a)\right)^2 \\
&= \sum_{a\in\mathcal{A}} G^{\mathrm{AM}}(a) - 4\sum_{a\in\mathcal{A}}\left[F^{\mathrm{AM}}(a)\right]^2 \\
&= \sum_{a\in\mathcal{A}}\left(G^{\mathrm{AM}}(a) - 4\left[F^{\mathrm{AM}}(a)\right]^2\right),
\end{aligned} \tag{121}$$

where $G^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$, $F^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + (1-f_a)^2 f_a^2$, and f_a is the probability of a minor allele occurring at locus a. Similar to the mean (Eq. 119), the variance just depends on minor allele probabilities f_a for all $a \in \mathcal{A}$.

With the mean (Eq. 119) and variance (Eq. 121) estimates of AM distances, the asymptotic AM distance distribution is given by the following

$$D_{ij}^{\mathrm{AM}} \dot{\sim} \mathcal{N}\left(2\sum_{a\in\mathcal{A}} F^{\mathrm{AM}}(a), \sum_{a\in\mathcal{A}}\left(G^{\mathrm{AM}}(a) - 4\left[F^{\mathrm{AM}}(a)\right]^2\right)\right), \tag{122}$$

where $G^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$, $F(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + (1-f_a)^2 f_a^2$, and f_a is the probability of a minor allele occurring at locus a.

This concludes our analysis of the AM metric in GWAS data when the independence assumption holds for minor allele probabilities f_a and binomial samples $\mathcal{B}(2, f_a)$ for all $a \in \mathcal{A}$. In the next section, we derive more complex asymptotic results for the TiTv distance metric (Eq. 112).

3.3 TiTv distance distribution

The TiTv metric allows for one to account for both genotype mismatch, allele mismatch, transition, and transversion. However, this added dimension of information requires

knowledge of the nucleotide makeup at a particular locus. A sufficient condition to compute the TiTv metric between instances $i, j \in \mathcal{I}$ 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). We illustrate all possibilities for transitions and transversions in a diagram (Fig. 2). 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 occurs only with PuPy encoding.

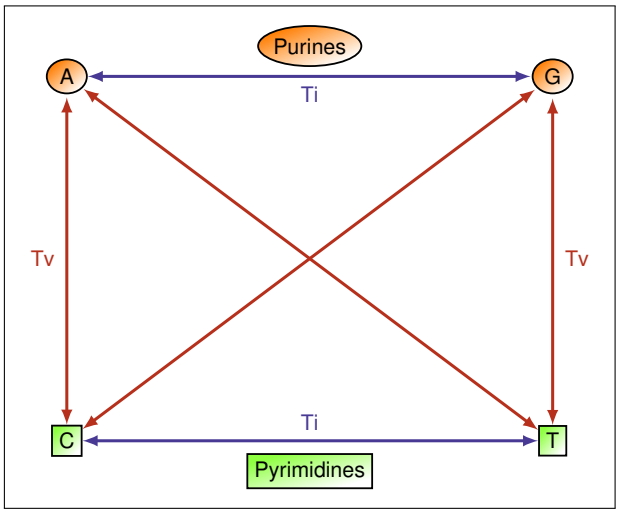

Fig 2. Purines (A and G) and pyrimidines (C and T) are shown. Transitions occur when a mutation involves purine-to-purine or pyrimidine-to-pyrimidine insertion. Transversions occur when a purine-to-pyrimidine or pyrimidine-to-purine insertion happens, which is a more extreme case. There are visibly more possibilities for transversions to occur than there are transitions, but there are about twice as many transitions in real data.

This additional encoding is always given in a particular GWAS data set, which leads us to consider the probabilities of PuPu, PuPy, and PyPy. These will be necessary to determine asymptotics for the TiTv distance metric. 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(Ti) is approximately twice the probability of transversion P(Tv). It is likely that any particular data set will not satisfy this criterion exactly. In this general case, we have P(Ti) being equal to some multiple η times P(Tv). In order to enforce this general constraint in simulated data, we define the following set of equalities

$$\gamma_0 + \gamma_1 + \gamma_2 = 1, \tag{123}$$
$$\mathrm{P(Ti)} - \eta \mathrm{P(Tv)} = 0. \tag{124}$$

The sum-to-one constraint (Eq. 123) is natural in this context because there are only three possible genotype encodings at a particular locus, which are PuPu, PuPy, and PyPy. Solving the Ti/Tv ratio constraint (Eq. 124) for η gives

$$\eta = \frac{\mathrm{P(Ti)}}{\mathrm{P(Tv)}},$$

which is easily computed in a real data set by dividing the fraction of Ti out of the total p loci by the fraction of Tv out of the total p loci. We will use the simplified notation $\eta = \mathrm{Ti/Tv}$ to represent this factor for the remainder of this work.

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

$$\mathrm{P(Tv)} = \gamma_1. \tag{125}$$

Using the sum-to-one constraint (Eqs. 123) and the probability of transversion (Eq. 124), the probability of a transition occuring at locus a is computed as follows

$$\mathrm{P}(\mathrm{Ti}) = \gamma_0 + \gamma_2. \tag{126}$$

Also using the sum-to-one constraint (Eq. 123) and the Ti/Tv ratio constraint (Eq. 124), it is clear that we have $\mathrm{P}(\mathrm{Tv}) = \frac{1}{\eta+1}$ and $\mathrm{P}(\mathrm{Ti}) = \frac{\eta}{\eta+1}$. Without loss of generality, we then sample

$$\gamma_0 \sim \mathcal{U}\left(\varepsilon, \frac{\eta}{\eta+1} - \varepsilon\right), \tag{127}$$

where ε is some small positive real number.

Then it immediately follows that we have

$$\gamma_2 = \frac{\eta}{\eta+1} - \gamma_0. \tag{128}$$

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 $\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = k\right]$ for each $k \in \mathcal{D} = \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}$. Let y represent a random sample of size p from $\{0, 1, 2\}$, where

$$y_a = \begin{cases} 0 & \text{if locus } a \text{ is PuPu,} \\ 1 & \text{if locus } a \text{ is PuPy,} \\ 2 & \text{if locus } a \text{ is PyPy.} \end{cases} \tag{129}$$

We derive $\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = 0\right]$ as follows

$$\begin{aligned} \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = 0\right] &= \mathrm{P}\left[y_a = 0, X_{ia} = X_{ja}\right] \\ &\quad + \mathrm{P}\left[y_a = 1, X_{ia} = X_{ja}\right] \\ &\quad + \mathrm{P}\left[y_a = 2, X_{ia} = X_{ja}\right] \\ &= \gamma_0\left[(1-f_a)^2 + 4f_a(1-f_a) + f_a^2\right] \\ &\quad + \gamma_1\left[(1-f_a)^2 + 4f_a(1-f_a) + f_a^2\right] \\ &\quad + \gamma_2\left[(1-f_a)^2 + 4f_a(1-f_a) + f_a^2\right] \\ &= (\gamma_0 + \gamma_1 + \gamma_2)\left[(1-f_a)^2 + 4f_a(1-f_a) + f_a^2\right] \\ &= (1-f_a)^2 + 4f_a(1-f_a) + f_a^2, \end{aligned} \tag{130}$$

where f_a is the probability of a minor allele occurring at locus a.

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

$$\begin{aligned} \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = \frac{1}{4}\right] &= 2\mathrm{P}\left[y_a = 0, X_{ia} = 0, X_{ja} = 1\right] \\ &\quad + 2\mathrm{P}\left[y_a = 0, X_{ia} = 1, X_{ja} = 2\right] \\ &\quad + 2\mathrm{P}\left[y_a = 2, X_{ia} = 0, X_{ja} = 1\right] \\ &\quad + 2\mathrm{P}\left[y_a = 2, X_{ia} = 1, X_{ja} = 2\right] \\ &= 4\gamma_0(1-f_a)^3 f_a + 4\gamma_0 f_a^3(1-f_a) + 4\gamma_2(1-f_a)^3 f_a \\ &\quad + 4\gamma_2 f_a^3(1-f_a) \\ &= 4\gamma_0\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] \\ &\quad + 4\gamma_2\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] \\ &= 4(\gamma_0 + \gamma_2)\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right], \end{aligned} \tag{131}$$

where f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu occurring at any locus a, and γ_2 is the probability of PyPy occurring at any locus a.

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

$$\begin{aligned}\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = \frac{1}{2}\right] &= 2\mathrm{P}\left[y_a = 1, X_{ia} = 0, X_{ja} = 1\right] \\ &\quad + 2\mathrm{P}\left[y_a = 1, X_{ia} = 1, X_{ja} = 2\right] \\ &= 4\gamma_1(1-f_a)^3 f_a + 4\gamma_1 f_a^3(1-f_a) \\ &= 4\gamma_1\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right],\end{aligned} \tag{132}$$

where f_a is the probability of a minor allele occurring at locus a and γ_1 is the probability of PuPy occurring at any locus a.

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

$$\begin{aligned}\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = \frac{3}{4}\right] &= 2\mathrm{P}\left[y_a = 0, X_{ia} = 0, X_{ja} = 2\right] \\ &\quad + 2\mathrm{P}\left[y_a = 2, X_{ia} = 0, X_{ja} = 2\right] \\ &= 2\gamma_0(1-f_a)^2 f_a^2 + 2\gamma_2(1-f_a)^2 f_a^2 \\ &= 2(\gamma_0+\gamma_2)(1-f_a)^2 f_a^2,\end{aligned} \tag{133}$$

where f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu occurring at any locus a, and γ_2 is the probability of PyPy occurring at any locus a.

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

$$\begin{aligned}\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = 1\right] &= 2\mathrm{P}\left[y_a = 1, X_{ia} = 0, X_{ja} = 2\right] \\ &= 2\gamma_1(1-f_a)^2 f_a^2,\end{aligned} \tag{134}$$

where f_a is the probability of a minor allele occurring at locus a and γ_1 is the probability of PuPy occurring at any locus a.

Using the TiTv diff probabilities (Eqs. 130-134), we compute the expected TiTv distance between instances $i, j \in \mathcal{I}$ as follows

$$\begin{aligned}\mathrm{E}\left(D_{ij}^{\mathrm{TiTv}}\right) &= \sum_{a\in\mathcal{A}}\left(\sum_{k\in\mathcal{D}} k\cdot\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = k\right]\right) \\ &= (\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}}\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] \\ &\quad + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}}(1-f_a)^2 f_a^2 \\ &= (\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}} F^{\mathrm{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\mathrm{TiTv}}(a),\end{aligned} \tag{135}$$

where $F^{\mathrm{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$, $G^{\mathrm{TiTv}}(a) = (1-f_a)^2 f_a^2$, f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu occurring at any locus a, γ_1 is the probability of PuPy occurring at any locus a, and γ_2 is the probability of PyPy occurring at any locus a. In contrast to the expected GM and AM distances (Eqs. 114 and 119), the expected TiTv distance (Eq. 135) depends on minor allele probabilities f_a for all $a \in \mathcal{A}$ and the genotype encoding probabilities $\gamma_0, \gamma_1,$ and γ_2.

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

$$
\begin{aligned}
\mathrm{E}\left[\left(D_{ij}^{\mathrm{TiTv}}\right)^2\right] &= \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\mathrm{TiTv}}(a)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a\in\mathcal{A}}\left(\mathrm{d}_{ij}^{\mathrm{TiTv}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r\in\mathcal{A}}\sum_{s\leq r-1} \mathrm{d}_{ij}^{\mathrm{TiTv}}(r)\cdot \mathrm{d}_{ij}^{\mathrm{TiTv}}(s)\right] \\
&= \sum_{a\in\mathcal{A}}\left(\sum_{k\in\mathcal{D}} k^2\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a)=k\right]\right) \\
&\quad + 2\sum_{a\in\mathcal{A}}\sum_{s\leq r-1}\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(r)=k\right]\right)\cdot\left(\sum_{k\in\mathcal{D}} k\cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(s)=k\right]\right) \\
&= \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F^{\mathrm{TiTv}}(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\mathrm{TiTv}}(a) \\
&+2\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}}\left([\gamma_0+\gamma_2+2\gamma_1]F^{\mathrm{TiTv}}(\lambda) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\mathrm{TiTv}}(\lambda)\right),
\end{aligned}
\tag{136}
$$

where $F^{\mathrm{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$, $G^{\mathrm{TiTv}}(a) = (1-f_a)^2 f_a^2$, f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu occurring at any locus a, γ_1 is the probability of PuPy occurring at any locus a, and γ_2 is the probability of PyPy occurring at any locus a.

Using the first (Eq. 135) and second (Eq. 136) raw moments of the TiTv distance, the variance is given by

$$
\begin{aligned}
\mathrm{Var}\left(D_{ij}^{\mathrm{TiTv}}\right) &= \mathrm{E}\left[\left(D_{ij}^{\mathrm{TiTv}}\right)^2\right] - \left[\mathrm{E}\left(D_{ij}^{\mathrm{TiTv}}\right)\right]^2 \\
&= \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F^{\mathrm{TiTv}}(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\mathrm{TiTv}}(a) \\
&+2\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}}\left([\gamma_0+\gamma_2+2\gamma_1]F^{\mathrm{TiTv}}(\lambda) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\mathrm{TiTv}}(\lambda)\right) \\
&\quad - \left([\gamma_0+\gamma_2+2\gamma_1]\sum_{a\in\mathcal{A}} F^{\mathrm{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\mathrm{TiTv}}(a)\right)^2 \\
&= \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F^{\mathrm{TiTv}}(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\mathrm{TiTv}}(a) \\
&\quad - \sum_{a\in\mathcal{A}}\left([\gamma_0+\gamma_2+2\gamma_1]F^{\mathrm{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\mathrm{TiTv}}(a)\right)^2,
\end{aligned}
\tag{137}
$$

where $F^{\mathrm{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$, $G^{\mathrm{TiTv}}(a) = (1-f_a)^2 f_a^2$, f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu occurring at any locus a, γ_1 is the probability of PuPy occurring at any locus a, and γ_2 is the probability of PyPy occurring at any locus a.

With the mean (Eq. 135) and variance (Eq. 137) estimates, the asymptotic TiTv

distance distribution is given by the following

$$D_{ij}^{\text{TiTv}} \dot{\sim} \mathcal{N}\Bigg((\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}} F^{\text{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\text{TiTv}}(a), \\ \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F^{\text{TiTv}}(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\text{TiTv}}(a) \\ -\sum_{a\in\mathcal{A}}\left([\gamma_0+\gamma_2+2\gamma_1]F^{\text{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\text{TiTv}}(a)\right)^2\Bigg), \tag{138}$$

where $F^{\text{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$, $G^{\text{TiTv}}(a) = (1-f_a)^2 f_a^2$, f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu occurring at any locus a, γ_1 is the probability of PuPy occurring at any locus a, and γ_2 is the probability of PyPy occurring at any locus a.

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

$$\bar{f}_a = \frac{1}{2}(l+u). \tag{139}$$

The maximum TiTv distance occurs at $\bar{f}_a = 0.5$ for any fixed Ti/Tv ratio η (Eq. 124), which is the inflection point about which the minor allele changes at locus a (Fig. 3). If few minor alleles are present ($\bar{f}_a \to 0$), the predicted TiTv distance approaches 0. The same is true after the minor allele switches ($\bar{f}_a \to 1$). To explore how TiTv distance changes with increased minor allele frequency, we fixed the Ti/Tv ratio η and generated simulated TiTv distances for $\bar{f}_a = 0.055, 0.150, 0.250,$ and 0.350 (Fig. 4A). For fixed η, TiTv distance increases significantly with increased $\bar{f}_a$. We similarly fixed the average minor allele frequency $\bar{f}_a$ and generated simulated TiTv distances for $\eta = \text{Ti/Tv} = 0.5, 1, 1.5,$ and 2 (Fig. 4C). The TiTv distance decreases slightly with increased $\eta = \text{Ti/Tv}$. As $\eta \to 0^+$, the data is approaching all Tv and no Ti, which means the TiTv distance is larger by definition. On the other hand, the TiTv distance decreases as $\eta \to 2^-$ because the data is approaching approximately twice as many Ti as there are Tv, which is typical for GWAS data in humans.

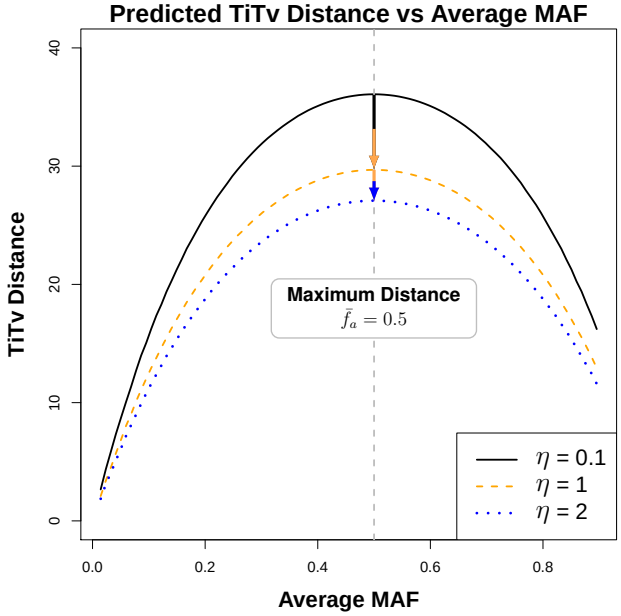

Fig 3. Predicted average TiTv distance as a function of average minor allele frequency $\bar{f}_a$ (see Eq. 139). Success probabilities f_a were drawn from a sliding window interval from 0.01 to 0.9 in increments of about 0.009. With $\eta = 0.1$, where η is the Ti/Tv ratio given by Eq. 123, Tv is ten 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.

We also compared theoretical and sample moments as a function of $\eta = \text{Ti/Tv}$ and $\bar{f}_a$ for the TiTv distance metric (Fig. 4B and D). We fixed $\bar{f}_a$ and computed the theoretical and simulated moments as a function of η (Fig. 4B). Theoretical average TiTv distance, given by Eq. 135, and simulated TiTv average distance are approximately equal as η increases. Theoretical standard deviation, given by Eq. 137, and simulated TiTv standard deviation differ slightly. We also fixed η and computed theoretical and sample moments as a function of $\bar{f}_a$ (Fig. 4D). In this case, there is approximate agreement with simulated and theoretical moments as $\bar{f}_a$ increases.

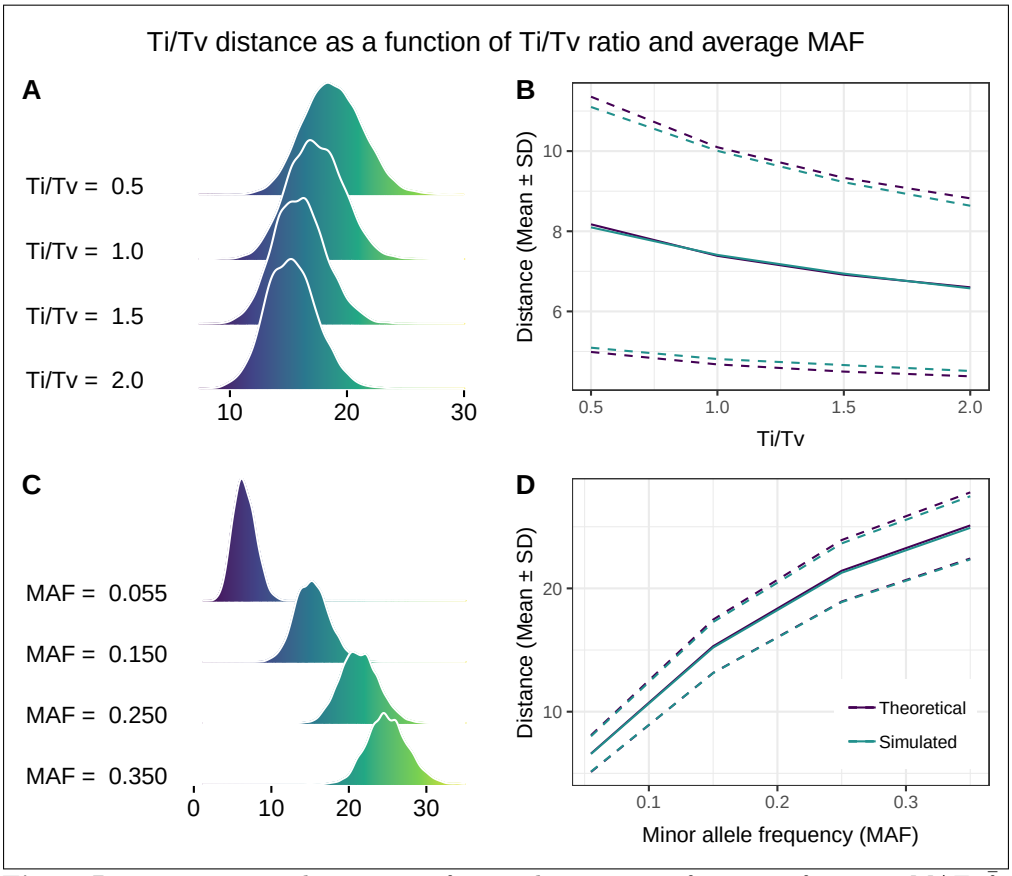

Fig 4. Density curves and moments of TiTv distance as a function of average MAF $\bar{f}_a$, given by Eq. 139, and Ti/Tv ratio η, given by Eq. 124. (**A**) For fixed $\bar{f}_a = 0.055$, TiTv distance density is plotted as a function of increasing η. TiTv distance decreases as η increases. For $\eta = \text{Ti/Tv} = 0.5$, there are twice as many transversions as there are transitions. On the other hand, $\eta = \text{Ti/Tv} = 2$ indicates that there are half as many transversions as transitions. Since transversions encode a larger magnitude distance than transitions, this behavior is expected. (**B**) Simulated and predicted mean ± SD are shown as a function of increasing Ti/Tv ratio η. Distance decreases as Ti/Tv increases. Theoretical and simulated moments are approximately the same. (**C**) For fixed $\eta = 2$, TiTv distance density is plotted as a function of increasing $\bar{f}_a$. TiTv distance increases as $\bar{f}_a$ approaches maximum of 0.5, which means that there is about the same frequency of minor alleles as major alleles. (**D**) Simulated and predicted mean ± SD as a function of increasing average MAF $\bar{f}_a$. Distance increases as the number of minor alleles increases. Theoretical and simulated moments are approximately the same.

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

GWAS-Metric	Stat	Formula (Eq. #)
GM (**Eq. 110**)	mean	$\boxed{2\sum_{a\in\mathcal{A}} F^{\mathrm{GM}}(a)}$ (**114**) where $F^{\mathrm{GM}}(a) = 2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2$
	variance	$\boxed{2\sum_{a\in\mathcal{A}} F^{\mathrm{GM}}(a)[1-2F^{\mathrm{GM}}(a)]}$ (**116**) where $F^{\mathrm{GM}}(a) = 2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2$
AM (**Eq. 111**)	mean	$\boxed{2\sum_{a\in\mathcal{A}} F^{\mathrm{AM}}(a)}$ (**119**) where $F^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + (1-f_a)^2 f_a^2$
	variance	$\boxed{\sum_{a\in\mathcal{A}} \left[G^{\mathrm{AM}}(a) - 4\left(F^{\mathrm{AM}}(a)\right)^2\right]}$ (**121**) where $F^{\mathrm{AM}}(a) = 2(1-f_a)^3 f_a + 2f_a^3(1-f_a) + (1-f_a)^2 f_a^2$ and $G^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$
TiTv (**Eq. 112**)	mean	$\boxed{(\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}} F^{\mathrm{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\mathrm{TiTv}}(a)}$ (**135**) where $F^{\mathrm{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G^{\mathrm{TiTv}}(a) = (1-f_a)^2 f_a^2$
	variance	$\boxed{\begin{aligned}&\left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}} F^{\mathrm{TiTv}}(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}} G^{\mathrm{TiTv}}(a)\\ &+\sum_{a\in\mathcal{A}}\left([\gamma_0+\gamma_2+2\gamma_1]F^{\mathrm{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\mathrm{TiTv}}(a)\right)^2\end{aligned}}$ (**137**) where $F^{\mathrm{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G^{\mathrm{TiTv}}(a) = (1-f_a)^2 f_a^2$

3.4 Distribution of one-dimensional projection of GWAS distance onto a SNP

We previously derived the exact distribution of the one-dimensional projected distance onto an attribute in continuous data (Sec. 2.2.3), which is used as the predictor in NPDR to calculate relative feature importance in the form of standardized beta coefficients. GWAS data and the metrics we have considered are discrete. Therefore, we derive the density function for each diff metric (Eqs. 107-109), which also serves as the probability distribution for each metric, respectively.

The support of the GM metric (Eq. 107) is simply $\{0,1\}$, so we derive the probability, $\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = k\right]$, of this diff taking on each of these two possible values. First, the probability that the GM diff is equal to zero is given by

$$\begin{aligned} f_{\mathrm{GM}}(0; f_a) = \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 0\right] &= \mathrm{P}\left(X_{ia}=0, X_{ja}=0\right) + \mathrm{P}\left(X_{ia}=1, X_{ja}=1\right) \\ &\quad + \mathrm{P}\left(X_{ia}=2, X_{ja}=2\right) \\ &= (1-f_a)^4 + 4f_a^2(1-f_a)^2 + f_a^4, \end{aligned} \tag{140}$$

where f_a is the probability of a minor allele occurring at locus a.

Similarly, the probability that the GM diff is equal to 1 is derived as follows

$$\begin{aligned} f_{\mathrm{GM}}(1; f_a) = \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 1\right] &= 2\mathrm{P}\left(X_{ia} = 0, X_{ja} = 1\right) + 2\mathrm{P}\left(X_{ia} = 1, X_{ia} = 2\right) \\ &\quad + 2\mathrm{P}\left(X_{ia} = 0, X_{ja} = 2\right) \\ &= 4\left(1 - f_a\right)^3 f_a + 4f_a^3\left(1 - f_a\right) + 2f_a^2\left(1 - f_a\right)^2, \end{aligned} \tag{141}$$

where f_a is the probability of a minor allele occurring at locus a.

This leads us to the probability distribution of the GM diff metric, which is the distribution of the one-dimensional GM distance projected onto a single SNP. This distribution is given by

$$f_{\mathrm{GM}}(d; f_a) = \begin{cases} \left(1 - f_a\right)^4 + 4f_a^2\left(1 - f_a\right)^2 + f_a^4 & d = 0, \\ 4\left(1 - f_a\right)^3 f_a + 4f_a^3\left(1 - f_a\right) + 2f_a^2\left(1 - f_a\right)^2 & d = 1, \end{cases} \tag{142}$$

where f_a is the probability of a minor allele occurring at locus a.

The mean and variance of this GM diff distribution can easily be derived using this newly determined density function (Eq. 142). The average GM diff is given by the following

$$\mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right] = 2F^{\mathrm{GM}}(a), \tag{143}$$

where $F^{\mathrm{GM}} = 2\left(1 - f_a\right)^3 f_a + 2f_a^3\left(1 - f_a\right) + f_a^2\left(1 - f_a\right)^2$ and f_a is the probability of a minor allele occurring at locus a.

The variance of the GM diff metric is given by

$$\mathrm{Var}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right] = 2F^{\mathrm{GM}}(a)\left[1 - 2F^{\mathrm{GM}}(a)\right], \tag{144}$$

where $F^{\mathrm{GM}} = 2\left(1 - f_a\right)^3 f_a + 2f_a^3\left(1 - f_a\right) + f_a^2\left(1 - f_a\right)^2$ and f_a is the probability of a minor allele occurring at locus a.

The support of the AM metric (Eq. 108) is $\{0, 1/2, 1\}$. Beginning with the probability of the AM diff being equal to 0, we have the following probability

$$\begin{aligned} f_{\mathrm{AM}}(0; f_a) = \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a) = 0\right] &= \mathrm{P}\left(X_{ia} = 0, X_{ja} = 0\right) + \mathrm{P}\left(X_{ia} = 1, X_{ja} = 1\right) \\ &\quad + \mathrm{P}\left(X_{ia} = 2, X_{ja} = 2\right) \\ &= \left(1 - f_a\right)^4 + 4f_a^2\left(1 - f_a\right)^2 + f_a^4, \end{aligned} \tag{145}$$

where f_a is the probability of a minor allele occurring at locus a.

The probability of the AM diff metric being equal to $1/2$ is computed similarly as follows

$$\begin{aligned} f_{\mathrm{AM}}\left(1/2; f_a\right) = \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a) = 1/2\right] &= 2\mathrm{P}\left(X_{ia} = 0, X_{ja} = 1\right) + 2\mathrm{P}\left(X_{ia} = 1, X_{ia} = 2\right) \\ &= 4\left(1 - f_a\right)^3 f_a + 4f_a^3\left(1 - f_a\right), \end{aligned} \tag{146}$$

where f_a the probability of a minor allele occurring at locus a.

Finally, the probability of the AM diff metric being equal to 1 is given by the following

$$\begin{aligned} f_{\mathrm{AM}}(1; f_a) = \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a) = 1\right] &= 2\mathrm{P}\left(X_{ia} = 0, X_{ja} = 2\right) \\ &= 2f_a^2\left(1 - f_a\right)^2, \end{aligned} \tag{147}$$

where f_a is the probability of a minor allele occuring at locus a.

As in the case of the GM diff metric, we now have the probability distribution of the AM diff metric. This also serves as the distribution of the one-dimensional AM distance projected onto a single SNP, and is given by the following

$$f_{\mathrm{AM}}(d; f_a) = \begin{cases} (1-f_a)^4 + 4f_a^2(1-f_a)^2 + f_a^4 & d = 0, \\ 4(1-f_a)^3 f_a + 4f_a^3(1-f_a) & d = 1/2, \\ 2f_a^2(1-f_a)^2 & d = 1, \end{cases} \tag{148}$$

where f_a is the probability of a minor allele occurring at locus a.

The mean and variance of this AM diff distribution is derived using the corresponding density function (Eq. 148). The average AM diff is given by

$$\mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a)\right] = 2F^{\mathrm{AM}}(a), \tag{149}$$

where $F^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + f_a^2(1-f_a)^2$ and f_a is the probability of a minor allele occurring at locus a.

The variance of the AM diff metric is given by

$$\mathrm{Var}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a)\right] = G^{\mathrm{AM}}(a) - 4\left[F^{\mathrm{AM}}(a)\right]^2, \tag{150}$$

where $G^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + 2(1-f_a)^2 f_a^2$, $F^{\mathrm{AM}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a) + f_a^2(1-f_a)$, f_a is the probability of a minor allele occurring at locus a.

For the TiTv diff metric (Eq. 109), the support is $\{0, 1/4, 1/2, 3/4, 1\}$. We have already derived the probability that the TiTv diff assumes each of the values of its support (Eqs. 130-134). Therefore, we have the following distribution of the TiTv diff metric

$$f_{\mathrm{TiTv}}(d; f_a, \gamma_0, \gamma_1, \gamma_2, \eta) = \begin{cases} (1-f_a)^4 + 4f_a^2(1-f_a)^2 + f_a^4 & d = 0, \\ 4(\gamma_0+\gamma_2)\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] & d = 1/4, \\ 4\gamma_1\left[(1-f_a)^3 f_a + f_a^3(1-f_a)\right] & d = 1/2, \\ 2(\gamma_0+\gamma_2)(1-f_a)^2 f_a^2 & d = 3/4, \\ 2\gamma_1(1-f_a)^2 f_a^2 & d = 1, \end{cases} \tag{151}$$

where f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu at locus a, γ_1 is the probability of PuPy at locus a, γ_2 is the probability of PyPy at locus a, and η is the Ti/Tv ratio (Eq. 124).

The mean and variance of this TiTv diff distribution is derived using the corresponding density function (Eq. 151). The average TiTv diff is given by

$$\mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a)\right] = (\gamma_0+\gamma_2+2\gamma_1)F^{\mathrm{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\mathrm{TiTv}}(a), \tag{152}$$

where $F^{\mathrm{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3(1-f_a)$, $G^{\mathrm{TiTv}}(a) = f_a^2(1-f_a)^2$, f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu at locus a, γ_1 is the probability of PuPy at locus a, and γ_2 is the probability of PyPy at locus a.

The variance of the TiTv diff metric is given by

$$\begin{aligned} \mathrm{Var}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a)\right] = {} & \left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]F^{\mathrm{TiTv}}(a) + \left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\mathrm{TiTv}}(a) \\ & - \left((\gamma_0+\gamma_2+2\gamma_1)F^{\mathrm{TiTv}}(a) + \left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G^{\mathrm{TiTv}}(a)\right)^2, \end{aligned} \tag{153}$$

where $F^{\text{TiTv}}(a) = (1-f_a)^3 f_a + f_a^3 (1-f_a)$, $G^{\text{TiTv}}(a) = f_a^2 (1-f_a)^2$, f_a is the probability of a minor allele occurring at locus a, γ_0 is the probability of PuPu at locus a, γ_1 is the probability of PuPy at locus a, and γ_2 is the probability of PyPy at locus a.

We now have the distribution of the projection of pairwise GWAS distances onto a single SNP. These distributions and their respective moments can inform NPDR, as well as other nearest-neighbor distance-based feature selection algorithms. Similar to the distribution of the TiTv pairwise distance, the TiTv diff distribution we have just derived is a novel result. In the next section, we introduce our new diff metric for time series correlation-based data, with a particular application to resting-state fMRI.

4 Time series correlation-based distance distribution

For time series correlation-based data, we consider the case where there are m correlation matrices $A^{(p\times p)}$. In particular, we are focusing on resting-state fMRI (rs-fMRI) data, which falls into this category. The derivations that follow, however, are relevant to all correlation-based data fitting the assumptions we have adopted. The features in rs-fMRI are commonly Regions of Interest (ROIs), which are collections of highly correlated and spatially proximal voxels [20]. These correlations are between different ROIs for a particular brain atlas [21]. Because the features are the ROIs themselves, this leads us to the following metric

$$\mathrm{d}_{ij}^{\text{ROI}}(a) = \sum_{k\neq a} \left| A_{ka}^{(i)} - A_{ka}^{(j)} \right|. \tag{154}$$

where $A_{ka}^{(i)}$ and $A_{ka}^{(j)}$ are the correlations between ROI a and ROI k for instances i and j, respectively. With this rs-fMRI diff, the pairwise distance between two instances $i, j \in \mathcal{I}$ is computed as follows

$$D_{ij}^{\text{fMRI}} = \sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\text{ROI}}(a). \tag{155}$$

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}
\\
\mathsf{ROI}_1 \\
\\
\\
\hline
\\
\mathsf{ROI}_2 \\
\\
\\
\hline
\vdots \\
\hline
\\
\mathsf{ROI}_p \\
\\
\\
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
\boldsymbol{I_1} & \boldsymbol{I_2} & \boldsymbol{I_3} & \boldsymbol{I_4} & \cdots & \boldsymbol{I_m}
\end{array} \\
\begin{bmatrix}
\hat{\boldsymbol{A}}_{12}^{(1)} & \hat{\boldsymbol{A}}_{12}^{(2)} & \hat{\boldsymbol{A}}_{12}^{(3)} & \hat{\boldsymbol{A}}_{12}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{12}^{(m)} \\
\hat{\boldsymbol{A}}_{13}^{(1)} & \hat{\boldsymbol{A}}_{13}^{(2)} & \hat{\boldsymbol{A}}_{13}^{(3)} & \hat{\boldsymbol{A}}_{13}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{13}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{\boldsymbol{A}}_{1p}^{(1)} & \hat{\boldsymbol{A}}_{1p}^{(2)} & \hat{\boldsymbol{A}}_{1p}^{(3)} & \hat{\boldsymbol{A}}_{1p}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{1p}^{(m)} \\
\hat{\boldsymbol{A}}_{21}^{(1)} & \hat{\boldsymbol{A}}_{21}^{(2)} & \hat{\boldsymbol{A}}_{21}^{(3)} & \hat{\boldsymbol{A}}_{21}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{21}^{(m)} \\
\hat{\boldsymbol{A}}_{23}^{(1)} & \hat{\boldsymbol{A}}_{23}^{(2)} & \hat{\boldsymbol{A}}_{23}^{(3)} & \hat{\boldsymbol{A}}_{23}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{23}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{\boldsymbol{A}}_{2p}^{(1)} & \hat{\boldsymbol{A}}_{2p}^{(2)} & \hat{\boldsymbol{A}}_{2p}^{(3)} & \hat{\boldsymbol{A}}_{2p}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{2p}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{\boldsymbol{A}}_{p1}^{(1)} & \hat{\boldsymbol{A}}_{p1}^{(2)} & \hat{\boldsymbol{A}}_{p1}^{(3)} & \hat{\boldsymbol{A}}_{p1}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{p1}^{(m)} \\
\hat{\boldsymbol{A}}_{p2}^{(1)} & \hat{\boldsymbol{A}}_{p2}^{(2)} & \hat{\boldsymbol{A}}_{p2}^{(3)} & \hat{\boldsymbol{A}}_{p2}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{p2}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{\boldsymbol{A}}_{p,p-1}^{(1)} & \hat{\boldsymbol{A}}_{p,p-1}^{(2)} & \hat{\boldsymbol{A}}_{p,p-1}^{(3)} & \hat{\boldsymbol{A}}_{p,p-1}^{(4)} & \cdots & \hat{\boldsymbol{A}}_{p,p-1}^{(m)}
\end{bmatrix}
\end{array}
= \boldsymbol{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. 38 and 39, that the mean and variance of the Manhattan ($q = 1$) distance distribution for standard normal data are $\frac{2p}{\sqrt{\pi}}$ and $\frac{2(\pi-2)p}{\pi}$, respectively. This allows us to easily derive the expected pairwise distance between instances i and j in rs-fMRI data as follows

$$
\begin{aligned}
\mathrm{E}(D_{ij}^{\mathrm{fMRI}}) &= \mathrm{E}\left(\sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\mathrm{ROI}}(a)\right) \\
&= \mathrm{E}\left(\sum_{a\in\mathcal{A}}\sum_{k\neq a}\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{a\in\mathcal{A}}\sum_{k\neq a}\mathrm{E}\left(\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{a\in\mathcal{A}}\sum_{k\neq a}\frac{2}{\sqrt{\pi}} \\
&= \frac{2p(p-1)}{\sqrt{\pi}}.
\end{aligned}
\tag{156}
$$

We first derive the variance of the rs-fMRI distance by making an independence assumption with respect to the magnitude differences $|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|$ for all $k \neq a \in \mathcal{A}$. We

proceed as follows

$$\begin{aligned}
\mathrm{Var}(D_{ij}^{\mathrm{fMRI}}) &= \mathrm{Var}\left(\sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\mathrm{ROI}}(a)\right) \\
&= \mathrm{Var}\left(\sum_{a\in\mathcal{A}}\sum_{k\neq a}\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{a\in\mathcal{A}}\sum_{k\neq a}\mathrm{Var}\left(\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{a\in\mathcal{A}}\sum_{k\neq a}\frac{2(\pi-2)}{\pi} \\
&= \frac{2(\pi-2)(p-1)p}{\pi}.
\end{aligned} \tag{157}$$

The independence assumption we made to derive the variance of our rs-fMRI distance metric (Eq. 157) is not actually satisfied due to the fact that a single value of the diff (Eq. 154) includes the same fixed ROI a for each term in the sum for all $k \neq a$. Therefore, the linear application of the variance operator we have previously employed does not account for the additional cross-covariance that exists. Although the theoretical variance of the distance we computed for the rs-fMRI distance metric (Eq. 157) still reasonably approximates the sample variance, there is a slight discrepancy between our theoretical rs-fMRI distance metric variance (Eq. 157) and the sample variance. More precisely, the formula we have given for the variance (Eq. 157) under-approximates the sample variance of the rs-fMRI distance. Because of this discrepancy, we determine a corrected formula by assuming that there is dependence between the terms of the rs-fMRI diff. We start the derivation of our corrected formula by writing the variance as a two-part sum, where the first term in the sum involves the variance of the magnitude difference $|\hat{A}_{ka}^{(i)} - \hat{A}_{ka}^{(j)}|$ and then second term involves the cross-covariance of the rs-fMRI diff for

distinct pairwise ROI-ROI associations. This begins as follows

$$
\begin{aligned}
\mathrm{Var}(D_{ij}^{\mathrm{fMRI}}) &= \mathrm{Var}\left(\sum_{a\in\mathcal{A}}\sum_{k\neq a}\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|\right)\\
&= \sum_{a=1}^{p-1}\mathrm{Var}\left(\sum_{k=a+1}^{p}2\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|\right)\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|,\sum_{s=r+1}^{p}2\left|\hat{A}_{rs}^{(i)}-\hat{A}_{rs}^{(j)}\right|\right)\\
&= \sum_{a=1}^{p-1}\sum_{k=a+1}^{p}\mathrm{Var}\left(2\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|\right)\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|,\sum_{s=r+1}^{p}2\left|\hat{A}_{rs}^{(i)}-\hat{A}_{rs}^{(j)}\right|\right) \qquad (158)\\
&= \sum_{a=1}^{p-1}\sum_{k=a+1}^{p-1}\frac{4(\pi-2)}{\pi}\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|,\sum_{s=r+1}^{p}2\left|\hat{A}_{rs}^{(i)}-\hat{A}_{rs}^{(j)}\right|\right)\\
&= \frac{2p(\pi-2)(p-1)}{\pi}\\
&+ 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|,\sum_{s=r+1}^{p}2\left|\hat{A}_{rs}^{(i)}-\hat{A}_{rs}^{(j)}\right|\right).
\end{aligned}
$$

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

$$
S^2_{D_{ij}^{\mathrm{fMRI}}} - \frac{2p(\pi-2)(p-1)}{\pi} = \beta_1 p^2 + \beta_0 p. \qquad (159)
$$

The coefficient estimates found through least squares fitting are $\beta_0 = -\beta_1 \approx 0.08$. These estimates allow us to arrive at a functional form for the double sum in the right-hand side of Eq. 158 that is proportional to $\frac{2p(\pi-2)(p-1)}{\pi}$. That is, we have the following formula for approximating the double sum

$$
2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p}2\left|\hat{A}_{ak}^{(i)}-\hat{A}_{ak}^{(j)}\right|,\sum_{s=r+1}^{p}2\left|\hat{A}_{rs}^{(i)}-\hat{A}_{rs}^{(j)}\right|\right) = \frac{p(\pi-2)(p-1)}{4\pi}. \qquad (160)
$$

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

$$
\mathrm{Var}(D_{ij}^{\mathrm{fMRI}}) = \frac{9p(\pi-2)(p-1)}{4\pi}. \qquad (161)
$$

With the mean and variance estimates given by Eqs. 156 and 161, we have the following asymptotic distribution for rs-fMRI distances

$$
D_{ij}^{\mathrm{fMRI}} \dot{\sim} \mathcal{N}\left(\frac{2p(p-1)}{\sqrt{\pi}}, \frac{9p(\pi-2)(p-1)}{4\pi}\right). \qquad (162)
$$

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

$$D_{ij}^{\text{fMRI}*} = \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{\left| A_{ak}^{(i)} - A_{ak}^{(j)} \right|}{\max(a) - \min(a)}. \tag{163}$$

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

$$\mathrm{E}\left(X_a^{\max} - X_a^{\min}\right) = 2\mu_{\max}^{(1)}(m,p) = 2\left[\frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m(p-1)}\right)} - \Phi^{-1}\left(\frac{1}{m(p-1)}\right)\right]. \tag{164}$$

The variance can be esimated with the following

$$\mathrm{Var}\left(X_a^{\max} - X_a^{\min}\right) = \frac{\pi^2}{6\log[m(p-1)]}. \tag{165}$$

Let $\mu_{D_{ij}^{\text{fMRI}}}$ and $\sigma^2_{D_{ij}^{\text{fMRI}}}$ denote the mean and variance of the rs-fMRI distance distribution given by Eqs. 156 and 161. Using the formulas for the mean and variance of the max-min normalized distance distribution given in Eq. 89, we have the following asymptotic distribution for the max-min normalized rs-fMRI distances

$$D_{ij}^{\text{fMRI}*} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{\text{fMRI}}}}{2\mu_{\max}^{(1)}(m,p)}, \frac{6\sigma^2_{D_{ij}^{\text{fMRI}}}\log[m(p-1)]}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m,p)\right]^2\log[m(p-1)]}\right). \tag{166}$$

4.1 One-dimensional projection of rs-fMRI distance onto a single ROI

Just as in previous sections (Secs. 2.2.3 and 3.4), we now derive the distribution of our rs-fMRI diff metric (Eq. 154). Unlike what we have seen in previous sections, we do not derive the exact distribution for this diff metric. We have determined empirically that the rs-fMRI diff is approximately normal. Although the rs-fMRI diff is a sum of $p-1$ magnitude differences, the Classical Central Limit Theorem does not apply because of the dependencies that exist between the terms of the sum. Examination of histograms and quantile-quantile plots of simulated values of the rs-fMRI diff easily indicate that the normality assumption is safe. Therefore, we derive the mean and variance of the approximately normal distribution of the rs-fMRI diff. As we have seen previously, this normality assumption is reasonable even for small values of p.

The mean of the rs-fMRI diff is derived by fixing a single ROI a and considering all pairwise associations with other ROIs $k \neq \mathcal{A}$. This is done as follows

$$\begin{aligned}
\mathrm{E}\left[\mathrm{d}_{ij}^{\text{ROI}}(a)\right] &= \mathrm{E}\left(\sum_{k\neq a}\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{k\neq a}\mathrm{E}\left(\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{k\neq a}\frac{2}{\sqrt{\pi}} \\
&= \frac{2(p-1)}{\sqrt{\pi}},
\end{aligned} \tag{167}$$

where a is a single fixed ROI.

Considering the variance of the rs-fMRI diff metric, we have two estimates. The first estimate uses the variance operator in a linear fashion, while the second will simply be a direct implication of the corrected formula of the variance of rs-fMRI pairwise distances (Eq. 161). Our first estimate is derived as follows

$$\begin{aligned}
\operatorname{Var}\left[\mathrm{d}_{ij}^{\mathrm{ROI}}(a)\right] &= \operatorname{Var}\left(\sum_{k\neq a}\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{k\neq a}\operatorname{Var}\left(\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|\right) \\
&= \sum_{k\neq a}\frac{2(\pi-2)}{\pi} \\
&= \frac{2(\pi-2)(p-1)}{\pi},
\end{aligned} \tag{168}$$

where a is a single fixed ROI.

Using the corrected rs-fMRI distance variance formula (Eq. 161), our second estimate of the rs-fMRI diff variance is given directly by the following

$$\operatorname{Var}\left[\mathrm{d}_{ij}^{\mathrm{ROI}}(a)\right] = \frac{9(\pi-2)(p-1)}{4\pi}, \tag{169}$$

where a is a single fixed ROI.

Empirically, the first estimate (Eq. 168) of the variance of our rs-fMRI diff is closer to the sample variance than the second estimate (Eq. 169). This is due to fact that we are considering only a fixed ROI $a \in mathcalA$, so the cross-covariance between the magnitude differences $\left|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}\right|$ for different pairs of ROIs (a and $k \neq a$) is negligible here. When considering all ROIs $a \in \mathcal{A}$, these cross-covariances are no longer negligible. Using the first variance estimate (Eq. 168) and the estimate of the mean (Eq. 167), we have the following asymptotic distribution of the rs-fMRI diff

$$\mathrm{d}_{ij}^{\mathrm{ROI}}(a) \dot{\sim} \mathcal{N}\left(\frac{2(p-1)}{\sqrt{\pi}}, \frac{2(\pi-2)(p-1)}{\pi}\right), \tag{170}$$

where a is a single fixed ROI.

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

Substituting the non-normalized mean given by Eq. 156 into Eq. 166 for the mean of the max-min normalized rs-fMRI metric, we have the following

$$\begin{aligned}
\mathrm{E}\left(D_{ij}^{\mathrm{fMRI}^*}\right) &= \frac{\mu_{D_{ij}^{\mathrm{fMRI}}}}{2\mu_{\max}^{(1)}(m,p)} \\
&= \frac{p(p-1)}{\sqrt{\pi}\mu_{\max}^{(1)}(m,p)},
\end{aligned} \tag{171}$$

where $\mu_{\max}^{(1)}(m,p)$ is given in Eq. 164.

Similarly, the variance of $D_{ij}^{\text{fMRI}^*}$ is given by

$$\begin{aligned}\text{Var}\left(D_{ij}^{\text{fMRI}^*}\right) &= \frac{6\sigma^2_{D_{ij}^{\text{fMRI}}}\log[m(p-1)]}{\pi^2 + 24\left[\mu^{(1)}_{\max}(m,p)\right]^2\log[m(p-1)]} \\ &= \frac{27(\pi-2)\log[m(p-1)](p-1)p}{2\pi\left(\pi^2 + 24\left[\mu^{(1)}_{\max}(m,p)\right]^2\log[m(p-1)]\right)},\end{aligned} \tag{172}$$

where $\mu^{(1)}_{\max}(m,p)$ is given in Eq. 164.

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

rs-fMRI - Metric	**Stat**	**Formula (Eq. #)**
standard (**Eq. 155**)	mean	$\frac{2p(p-1)}{\sqrt{\pi}}$ (**156**)
	variance	$\frac{9p(\pi-2)(p-1)}{4\pi}$ (**158**)
max-min normalized (**Eq. 163**)	mean	$\boxed{\frac{\mu_{D_{ij}}}{2\mu^{(1)}_{\max}(m,p)}}$ (**171**) where $\mu_{D_{ij}}$ and $\mu^{(1)}_{\max}(m,p)$ are given by Eqs. 156 and 164
	variance	$\boxed{\frac{6\sigma^2_{D_{ij}}\log[m(p-1)]}{\pi^2 + 24\left[\mu^{(1)}_{\max}(m,p)\right]^2\log[m(p-1)]}}$ (**172**) where $\sigma^2_{D_{ij}}$ and $\mu^{(1)}_{\max}(m,p)$ are given by Eqs. 156 and 164

5 Comparison of theoretical and sample moments

It is important that our asymptotic estimates of sample moments for distances in high-dimensional data are accurate and precise enough to be useful when considering either simulated or real data that satisfy the appropriate assumptions. We compared our asymptotic estimates of sample moments of pairwise distances by generating random data for various dimensions m and p (Fig. 6). We fixed $m = 100$ and generated random Manhattan (Eq. 1) distance matrices on standard normal data for $p = 1000, 2000, 3000, 4000,$ and 5000. For each value of p, we generated 20 random distance matrices, computed the average distance, and computed the standard deviation of distances. We then computed the average of these 20 different means and standard deviations, respectively. The theoretical moments were simply computed for each value of p and fixed $m = 100$ in order to determine how closely our formulas approximate simulated moments. We generated scatter plots of theoretical vs simulated mean (Fig. 6A) and theoretical vs simulated standard deviation (Fig. 6B). All points lie approximately on the dashed identity lines, which implies that our asymptotic formulas for sample moments are reliable for both large and relatively small numbers of features. Although

we do not show these results for every combination of data type, distance metric, sample size m, and number of features p, all theoretical formulas of sample moments we have derived similarly approximate their corresponding simulated moments.

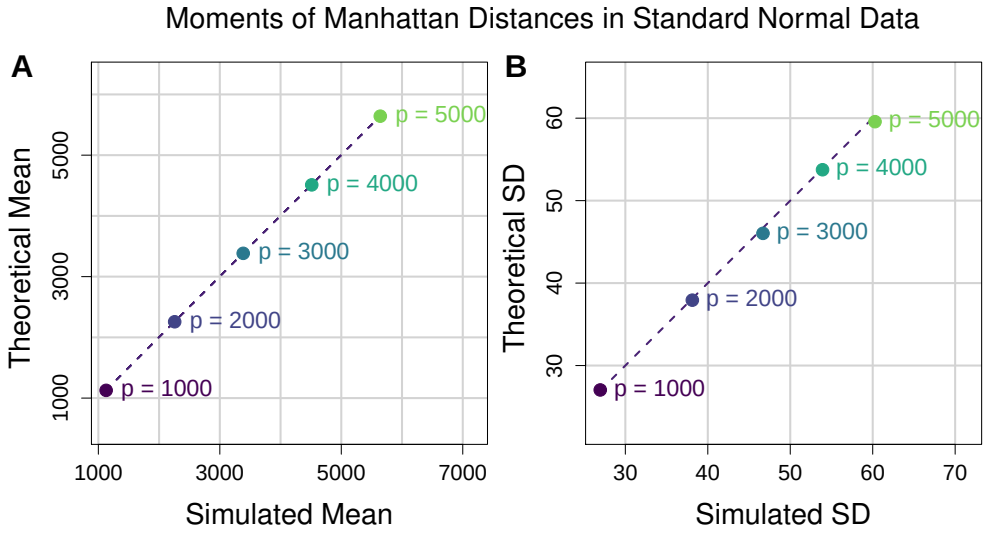

Fig 6. Comparison of theoretical and sample moments of Manhattan (Eq. 1) distances in standard normal data. (**A**) Scatter plot of theoretical vs simulated mean Manhattan distance (Eq. 38). Each point represents a different number of attributes p. For each value of p we fixed $m = 100$ and generated 20 distance matrices from standard normal data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of p for comparison. The dashed line represents the identity (or $y = x$) line for reference. (**B**) Scatter plot of theoretical vs simulated standard deviation of Manhattan (Eq. 1) distance (Eq. 39). These standard deviations come from the same random distance matrices for which mean distance was computed for **A**. Both theoretical mean and standard deviation approximate the simulated moments quite well.

6 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.

6.1 Continuous data

Without loss of generality, suppose that we have $X^{(m\times p)}$ where $X_{ia} \sim \mathcal{N}(0,1)$ for all $i = 1, 2, \ldots, m$ and $a = 1, 2, \ldots, p$, and let $m = p = 100$ and consider only the L_2 (Euclidean) metric (Eq. 1, $q = 2$). We explore the effects of correlation on these distances by generating simulated data sets with increasing degree of pairwise feature correlation and then plotting the density curve of the induced distances (Fig. 7A). 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

$$\bar{r}_{\text{abs}} = \frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j>i} r_{ij} \tag{173}$$

where r_{ij} is the correlation between features $i, j \in \mathcal{I}$ across all instances m. Distances generated on data without correlation closely approximate a Gaussian. The mean and variance of the uncorrelated distance distribution are given by Eqs. 50 and 49, respectively, by substituting $p = 100$ for the mean. As $\bar{r}_{\text{abs}}$ increases, we very quickly see positive skewness and increased variability in distances. The predicted and sample means, however, are approximately the same between correlated and uncorrelated distances due to linearity of the expectation operator. Because of the dependencies between features, the predicted variance of 1 for L_2 on standard normal data obviously no longer holds.

In order to introduce a controlled level of correlation between features, we created correlation matrices based on a random graph with specified connection probability, where features correspond to the vertices in each graph. We assigned high correlations to connected features from the random graph and low correlations to all non-connections. Using the upper-triangular cholesky factor U for uncorrelated data matrix X, we computed the following product to create correlated data matrix X^{corr}

$$X^{\text{corr}} = XU^{\text{T}}. \tag{174}$$

The new data matrix given by Eq. 174 has approximately the same correlation structure as the randomly generated correlation matrix created from a random graph. The cholesky method is a standard approach in creating correlated data sets.

6.2 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 only the TiTv metric, which is given by combining Eqs. 109 and 1 with $q = 1$. To create correlated GWAS data, we first generated standard normal data with random correlation structure, just as in the previous section. We then applied the standard normal cumulative distribution function (CDF) to this correlated data in order transform the correlated standard normal variates into uniform data with preserved correlation structure. We then subsequently applied the inverse binomial CDF to the correlated uniform data with random success probabilities f_a for all $a \in \mathcal{A}$. Each feature $a \in \mathcal{A}$ corresponds to an individual SNP in the data matrix. The resulting GWAS data set is binomial with $n = 2$ trials and has roughly the same correlation matrix as the original correlated standard normal data with which we started. Average absolute pairwise correlation $\bar{r}_{\text{abs}}$ induces positive skewness in GWAS data at lower levels than in correlated standard normal data (Fig. 7B). This could have important implications in nearest neighborhoods in NPDR and other similar methods.

6.3 Correlation-based data

For our correlation data-based metric given by Eqs. 154 and 1 with $q = 1$, we consider additional effects of correlation between features. Without loss of generality, we let $m = 100$ and $p = 30$. We show an illustration of the effects of correlated features in this context (Fig. 7C). Based on the correlated distance densities, it appears that correlation between features introduces positive skewness at lower values of $\bar{r}_{\text{abs}}$. We introduced correlation to the transformed data matrix (Fig. 5) with the cholesky method used previously.

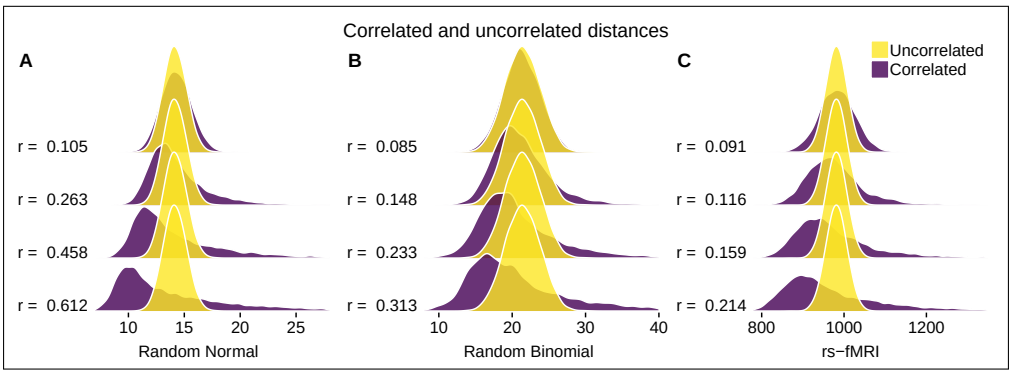

Fig 7. Ridgeline plots of distance densities from uncorrelated vs correlated bioinformatics data. (**A**) Euclidean distance densities for random normal data with and without correlation. Correlated data was created by multiplying random normal data by upper-triangular cholesky factor from randomly generated correlation matrix. We created correlated data for average absolute pairwise correlation (Eq. 173) $\bar{r}_{\text{abs}} = 0.105, 0.263, 0.458,$ and 0.612. (**B**) TiTv distance densities for random binomial data with and without correlation. Correlated data was created by first generating correlated standard normal data using the cholesky method from (A). Then we applied the standard normal CDF to create correlated uniformly distributed data, which was then transformed by the inverse binomial CDF with $n = 2$ trials and success probabilites f_a for all $a \in \mathcal{A}$. The result is correlated random binomial data with correlation matrix approximately the same as the original correlated standard normal data. (**C**) Time series correlation-based distance densities for random rs-fMRI data (Fig. 5) with and without additional pairwise feature correlation. Correlation was added to the transformed rs-fMRI data matrix (Fig. 5) using the cholesky algorithm from (A).

7 Discussion

Nearest-neighbor distance-based feature selection is class of methods that are relatively simple to implement, intuitive in nature, and perform surprisingly well in detecting interaction effects in high dimensional data. However, there has been little work done to understand how the limiting behavior of distance distributions can aid in determining how to properly parameterize these methods for feature selection. Furthermore, little has been done in the way of optimizing the choice of distance metric. Most often, distance-based feature selection methods use the L_q metric given by Eq. 1 with $q = 1$ or $q = 2$. However, these two realizations of the L_q metric have considerably different expressions for the mean and variance of their respective limiting distributions. For instance, the expected distance for L_1 and L_2 on standard normal data is on the order of p (see Eq. 38) and $\sqrt{p}$ (see Eq. 48), respectively. In addition, L_1 and L_2 on standard normal data have asymptotic variances on the order of p and 1, respectively. Considering whether one should choose L_1 or L_2 in this context may depend on motivation. For

instance, distances become harder to distinguish from one another in high dimensions, which is one of the curses of dimensionality. In the case of L_2, unit variance in the limit distribution means that distances will be almost completely contained within a ball of radius 1. The limiting L_2 distribution can therefore be thought of simply as a positive translation of the standard normal distribution. On the other hand, the L_1 distances become more dispersed due to the fact that the variance of the limiting distribution is proportional to the feature dimension p. This could actually be more desirable when determining nearest neighbors because instances may be easier to distinguish with this metric. If using L_1, then it may be best to use a fixed-k algorithm instead of fixed-radius. This is because fixed-radius neighborhood order could vary quite a bit considering the L_1 variance is proportional to feature dimension p, which in turn could affect the quality of selected features. If L_2 is being used, then perhaps either fixed-k or fixed-radius may perform equally well because most distances will be within 1 standard deviation away from the mean.

In any neighborhood selection algorithm, it is important to know what the average distance is and how dispersed these distances become as the feature dimension p grows. In our analysis, we have derived distance asymptotics for some of the most commonly used metrics in nearest-neighbor distance-based feature selection, as well as two new metrics for GWAS and time series correlation-based data like resting-state fMRI. Using extreme value theory, we have derived limiting distributions for the sample maximum and minimum of a fixed feature a. This has allowed us to determine the expected value and variance of the max-min normalized L_q distance in standard normal and standard uniform data, which is a new result to the best of our knowledge. Our derivations provide an important reference for individuals that are using nearest-neighbor feature selection methods in typical bioinformatics data.

In this work, we have expanded nearest-neighbor distance-based feature selection into the context of time series correlation-based data. Our motivation for this is partly based on the fact that these methods have not yet been applied to resting-state fMRI data. In order for this to be possible, we had to create a metric (see Eq. 154) that could allow us to have regions of interest (ROIs) as features. Not all ROIs will be relevant to a particular phenotype in case-control studies, so it could be important to use a nearest-neighbor feature selection method to determine which ROIs are important. This could allow us to detect interactions to help elucidate the network structure of the brain as it relates to the phenotype of interest.

The recently introduced transition-transversion metric given by Eq. 109 provides an additional dimension to the commonly used discrete metrics in GWAS nearest-neighbor distance-based feature selection. In this work, we have provided the asymptotic mean and variance of the limiting TiTv distance distribution. This novel result, as well as asymptotic estimates for the GM (see Eq. 107) and AM (see Eq. 108) metrics, provides an important reference to aid in neighborhood parameter selection in this context. We have also shown how the Ti/TV ratio η (see Eq. 124) and minor allele frequency (or success probability) f_a affects these discrete distances. For the GM and AM metrics, the distance is solely determined by the minor allele frequencies because the genotype encoding is not taken into account. We showed how both minor allele frequency and Ti/Tv ratio uniquely affects the TiTv distance (Figs. 4A and 4C). Because transversions are more drastic forms of mutation than transitions, this additional dimension of information is important to consider, which is why we have provided asymptotic results for this metric.

Correlations exist between features and instances in real data. Because of this, there can be rather drastic divergence from the asymptotic results for uncorrelated data we have derived in this work. To illustrate this behavior, we showed how strong correlations lead to positive skewness in the distance distribution of random normal, binomial, and rs-fMRI data (Figs. 7A, 7B, and 7C). Pairwise correlation between features does not

change the average distance, so our asymptotic results for uncorrelated data also apply when features are not independent. In contrast, the sample variance of distances diverges from the uncorrelated case substantially as the average absolute pairwise feature-feature correlation increases (see Eq. 173). For fixed-radius neighborhood methods, this increases the probability of neighborhood inclusion for a fixed instance. The increased variability in distances on correlated data may provide further motivation for optimizing the choice of metric in nearest-neighbor feature selection. This most certainly motivates a discussion on optimal choices of neighborhood selection parameters, which we will address in future work.

There are many different distance metrics that can be used in place of those we have considered for bioinformatics data, but we have derived results for those that are the most commonly used in practice. Our work brings together many important aspects of nearest-neighbor distance-based feature selection, which also serves as a guide to other researchers that may be interested in a different choice of metric for a similar analysis. In future work, we will consider how pairwise feature correlation, as well as a mixture of main and interaction effects, changes the optimal choice of neighborhood selection parameters like fixed-k and fixed-radius.

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

5. 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.

6. 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.

7. 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.

8. 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.

9. Stephen J. Gotts, W. Kyle Simmons, Lydia A. Milbury, Gregory L. Wallace, Robert W. Cox, and Alex Martin. Fractionation of social brain circuits in autism spectrum disorders. *Brain*, 135:2711–2725, 2012.

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

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

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

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

14. Alvis Brazma and Jaak Vilo. Gene expression data analysis. *FEBS Letters*, 480:17–24, June 2000.

15. Dongfang Wang and Jin Gu. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genomics Proteomics Bioinformatics*, 16:320–331, December 2018.

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

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

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

19. Pei Li, Maozu Guo, Chunyu Wang, Xiaoyan Liu, and Quan Zou. An overview of SNP interactions in genome-wide association studies. *Briefings in Functional Genomics*, 14(2):143–155, September 2014.

20. Megan H. Lee, Christopher D. Smyser, and Joshua S. Shimony. Resting state fMRI: A review of methods and clinical applications. *AJNR Am J Neuroradiol.*, 34(10):1866–1872, October 2013.

21. David Alexander Dickie, Susan D. Shenkin, Devasuda Anblagan, Juyoung Lee, Manuel Blesa Cabez, David Rodriguez, James P. Boardman, Adam Waldman, Dominic E. Job, and Joanna M. Wardlaw. Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging. *Frontiers in Neuroinformatics*, January 2017.