Novel metrics and nearest-neighbor distance distributions in 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

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

Author summary

Introduction

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

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

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

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

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

In Section 1, we introduce some 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 applicable to continuously distributed data sets with m instances and p features. To illustrate specific realizations of these general derivations, we focus on the cases of standard normal and standard uniform data distributions. We address max-min normalized versions of standard L_q metrics using Extreme Value Theory (EVT) in Section 2.4. In Section 3, we then make a transition to discrete data in which each value in the $m \times p$ data matrix is from a binomial distribution parameterized by $n = 2$ trials and some success probability. 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, we show how correlation in the attribute space changes distance distributional properties in Section 5.

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 $d_{ij}(a)$, known as a "diff" in Relief literature, is the projection of the distance between instances i and j onto the attribute a dimension. The function $d_{ij}(a)$ supports any type of attributes (e.g., numeric and categorical). For example, the projected difference between two instances i and j for a continuous numeric (d^{num}) attribute a may be

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

where $\hat{X}$ represents the standardized data matrix X. We use a simplified $d_{ij}(a)$ notation in place of the $\text{diff}(a, (i, j))$ 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 $d_{ij}^{num}(a)$ 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 $|X_{ia} - X_{ja}|^q = |d_{ij}(a)|^q$ is another random variable. Let $Z_a^q \sim \mathcal{F}_{Z^q}(\mu_{z^q}, \sigma_{z^q}^2)$ be the random variable such that

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

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

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

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 [10] can be applied to show that

$$g\left(\left(D_{ij}^{(q)}\right)^q\right) = g\left(\sum_{a\in\mathcal{A}}^{p} Z_a^q\right)$$

$$= \left(\sum_{a\in\mathcal{A}} |X_{ia} - X_{ja}|^q\right)^{1/q} \tag{5}$$

$$= D_{ij}^{(q)} \overset{\cdot}{\sim} \mathcal{N}\left(g\left(\mu_{z^q}p\right), \left[g\prime\left(\mu_{z^q}p\right)\right]^2 \sigma_{z^q}^2 p\right)$$

$$\Rightarrow \; D_{ij}^{(q)} \overset{\cdot}{\sim} \mathcal{N}\left((\mu_{z^q}p)^{1/q}, \frac{\sigma_{z^q}^2 p}{q^2\left(\mu_{z^q}p\right)^{2\left(1-\frac{1}{q}\right)}}\right).$$

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

$$\mathrm{E}\left(D_{ij}^{(2)}\right) = \sqrt{\mathrm{E}\left[\left(D_{ij}^{(2)}\right)^2\right] - \mathrm{Var}\left(D_{ij}^{(2)}\right)}$$

$$= \sqrt{\mu_{z^2}p - \frac{\sigma_{z^2}^2}{4\mu_{z^2}}}. \tag{6}$$

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 (Figs. 1 and 2). We compute the distance between all pairs of instances in simulated datasets of uniformly and normally 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, Figs. 1 and 2 have an analogous representation for discrete data.
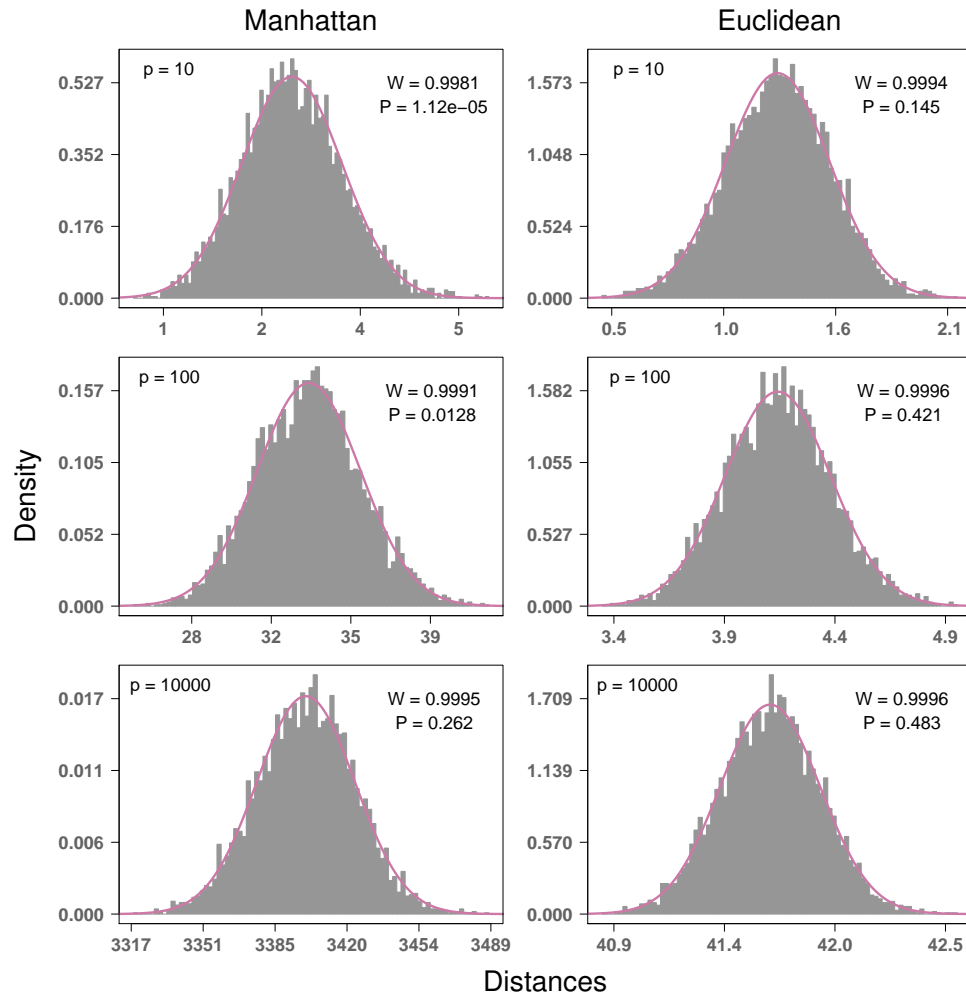
Fig 1. Convergence to Gaussian for Manhattan and Euclidean distances for simulated standard uniform data with $m = 100$ instances and $p = 10, 100,$ and 10000 attributes. Convergence to Gaussian occurs rapidly with increasing p, and Gaussian is a good approximation for p as low as 10 attributes. The 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.
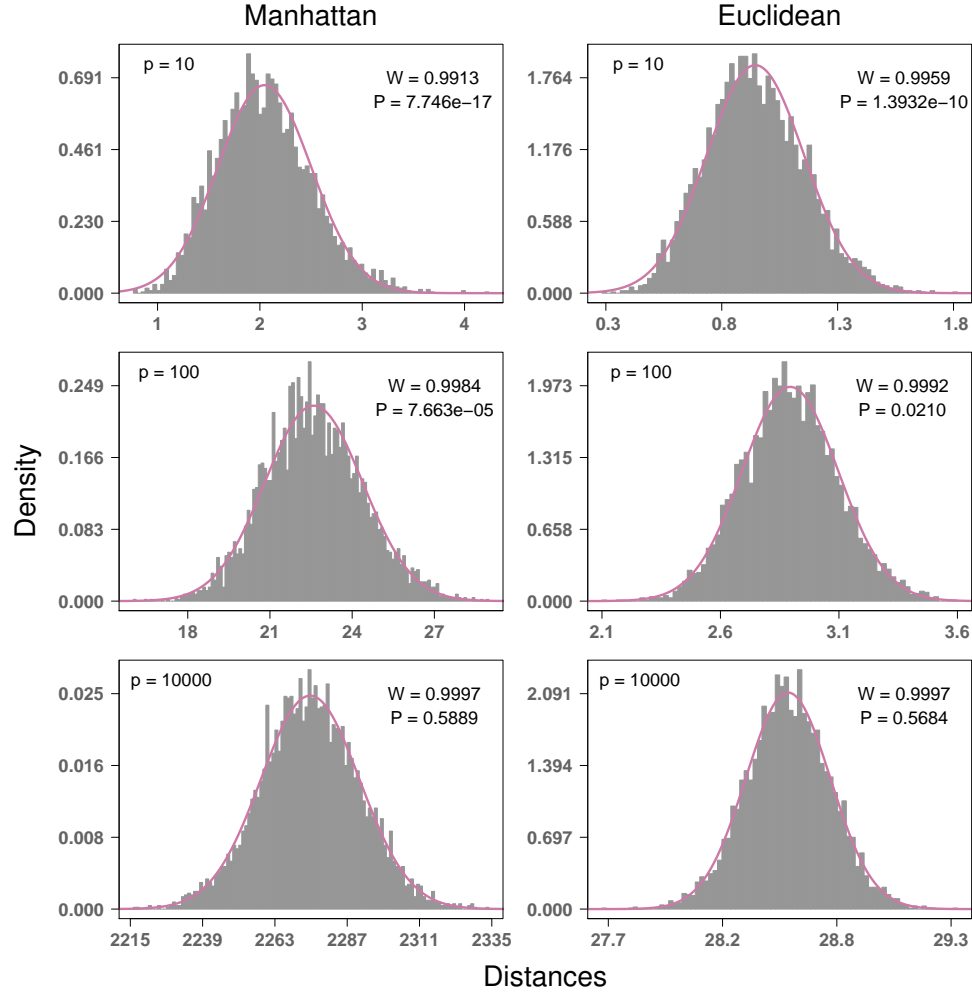
Fig 2. Convergence to Gaussian for Manhattan and Euclidean distances for simulated standard normal data with $m = 100$ instances and $p = 10, 100,$ and 10000 attributes. Convergence to Gaussian occurs rapidly with increasing p, and Gaussian is a good approximation for p as low as 10 attributes. The number 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 $Z_a^q = |\mathbf{d}_{ij}(a)|^q = |X_{ia} - X_{ja}|^q$, where $a \in \mathcal{A}$ and $|\mathcal{A}| = p$. In order to find the distribution of Z_a^q, we will use the following theorem given in [11].

Theorem 2.1 *Let $f(x)$ be the value of the probability density of the continuous random variable X at x. If the function given by $y = u(x)$ is differentiable and either increasing or decreasing for all values within the range of X for which $f(x) \neq 0$, then, for these values of x, the equation $y = u(x)$ can be uniquely solved for x to give $x = w(y)$, and for the corresponding values of y the probability density of $Y = u(X)$ is given by*

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

Elsewhere, $g(y) = 0$.

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

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

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

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

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

(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

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

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

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

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

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

and

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

Then it follows from fundamental rules of probability that

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

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

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

where $z_a > 0$.

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

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

and

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

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

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

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

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

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

$$f_{Z^q}(z_a^q) = \frac{1}{2q\pi (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \left(e^{-\frac{1}{2}\left[\sqrt{2}x_{ia} - \frac{\sqrt{2}}{2}(z_a^q)^{1/q}\right]^2}\right.$$
$$\left. + e^{-\frac{1}{2}\left[\sqrt{2}x_{ia} + \frac{\sqrt{2}}{2}(z_a^q)^{1/q}\right]^2}\right) \mathrm{d}x_{ia} \tag{22}$$

$$= \frac{1}{2q\sqrt{\pi} (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left(e^{-\frac{1}{2}u^2} + e^{-\frac{1}{2}u^2}\right) \mathrm{d}u \tag{23}$$

$$= \frac{1}{2q\sqrt{\pi} (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} (1+1) \tag{24}$$

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

$$= \frac{\frac{2}{q}}{(2^q)^{1/q} \Gamma\left(\frac{\frac{1}{q}}{\frac{2}{q}}\right)} (z_a^q)^{\frac{1}{q}-1} e^{-\left(\frac{z_a^q}{2^q}\right)^{2/q}}. \tag{26}$$

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

$$
\begin{aligned}
\mu_{z^q} &= \frac{c\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \\
&= \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}
\end{aligned}
\tag{27}
$$

and

$$
\begin{aligned}
\sigma_{z^q}^2 &= c^2\left[\frac{\Gamma\left(\frac{d+2}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} - \left(\frac{\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)}\right)^2\right] \\
&= 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right].
\end{aligned}
\tag{28}
$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 27 and 28 allow the p-dimensional mean and variance of the $D_{ij}^{(q)}$ distribution to be computed directly as

$$
\mu_{\left(D_{ij}^{(q)}\right)^q} = \mathrm{E}\left[\left(D_{ij}^{(q)}\right)^q\right] = \mathrm{E}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) = \sum_{a\in\mathcal{A}}\mathrm{E}\left(Z_a^q\right) = \sum_{a\in\mathcal{A}}\frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} = \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}p
\tag{29}
$$

and

$$
\begin{aligned}
\sigma_{\left(D_{ij}^{(q)}\right)^q}^2 &= \mathrm{Var}\left[\left(D_{ij}^{(q)}\right)^q\right] = \mathrm{Var}\left(\sum_{a\in\mathcal{A}} Z_a^q\right) \\
&= \sum_{a\in\mathcal{A}}\mathrm{Var}\left(Z_a^q\right) \\
&= \sum_{a\in\mathcal{A}} 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right] \\
&= 4^q\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]p.
\end{aligned}
\tag{30}
$$

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

$$
\mathcal{N}\left(\left(2^q\frac{\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/q}, \frac{4^q p}{q^2\left(\frac{2^q\Gamma\left(\frac{1}{2}q+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{\Gamma\left(q+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q+\frac{1}{2}\right)}{\pi}\right]\right).
\tag{31}
$$

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}, $X_{ia} - (Z_a^q)^{1/q}$, and $X_{ia} + (Z_a^q)^{1/q}$ are defined as

$$
f_X(x_{ia}) = 1, \quad 0 \le x_{ia} \le 1
\tag{32}
$$

$$
f_X\left(x_{ia} - (z_a^q)^{1/q}\right) = 1, \quad 0 \le x_{ia} - (z_a^q)^{1/q} \le 1, \text{ and}
\tag{33}
$$

$$f_X \left(x_{ia} + (z_a^q)^{1/q} \right) = 1, \quad 0 \le x_{ia} + (z_a^q)^{1/q} \le 1. \tag{34}$$

Substituting the results given by Eqs. 32-34 into Eq. 14, we have

$$f_{Z^q}(z_a^q) = \frac{1}{q \left(z_a^q \right)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[f_X \left(x_{ia} - (z_a^q)^{1/q} \right) + f_X \left(x_{ia} + (z_a^q)^{1/q} \right) \right] \mathrm{d}x_{ia},$$

$$0 < z_a \le 1$$

$$= \frac{1}{q \left(z_a^q \right)^{1-\frac{1}{q}}} \int_0^1 [f_X(x_{ia} - (z_a^q)) + f_X \left(x_{ia} + (z_a^q)^{1/q} \right)] \mathrm{d}x_{ia}, \quad 0 < z_a \le 1$$

$$= \frac{1}{q \left(z_a^q \right)^{1-\frac{1}{q}}} \int_{(z_a^q)}^1 1 \mathrm{d}x_{ia} + \int_0^{1-(z_a^q)} 1 \mathrm{d}x_{ia}, \quad 0 < z_a \le 1$$

$$= \frac{1}{q \left(z_a^q \right)^{1-\frac{1}{q}}} \left[(1 - (z_a^q)) + (1 - (z_a^q)) \right], \quad 0 < z_a \le 1$$

$$= \frac{1}{q} \cdot 2 \left(z_a^q \right)^{\frac{1}{q}-1} \left[1 - (z_a^q)^{1/q} \right]^{2-1}, \quad 0 < z_a \le 1. \tag{35}$$

The density given by Eq. 35 is a Kumaraswamy density with parameters $b = \frac{1}{q}$ and $c = 2$ with moment generating function (MGF) given by

$$\begin{aligned} M_n &= \frac{c\Gamma \left(1 + \frac{n}{b} \right) \Gamma(c)}{\Gamma \left(1 + c + \frac{n}{b} \right)} \\ &= \frac{2}{(nq+2)(nq+1)}. \end{aligned} \tag{36}$$

Using the MGF given by Eq. 36, the mean and variance of Z_a^q are computed as

$$\mu_{z^q} = \frac{2}{(q+2)(q+1)} \tag{37}$$

and

$$\sigma_{z^q}^2 = \frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)} \right)^2. \tag{38}$$

By linearity of the expected value and variance operators under the iid assumption, Eqs. 39 and 40 allow the p-dimensional mean and variance of the $\left(D_{ij}^{(q)} \right)^q$ distribution to be computed directly as

$$\begin{aligned} \mu_{\left(D_{ij}^{(q)} \right)^q} = \mathrm{E} \left[\left(D_{ij}^{(q)} \right)^q \right] &= \mathrm{E} \left(\sum_{a \in \mathcal{A}} Z_a^q \right) \\ &= \sum_{a \in \mathcal{A}} \mathrm{E}(Z_a^q) \\ &= \sum_{a \in \mathcal{A}} \frac{2}{(q+2)(q+1)} \\ &= \frac{2p}{(q+2)(q+1)} \end{aligned} \tag{39}$$

and

$$\sigma^2_{\left(D^{(q)}_{ij}\right)^q} = \mathrm{Var}\left[\left(D^{(q)}_{ij}\right)^q\right] = \mathrm{Var}\left(\sum_{a\in\mathcal{A}} Z^q_a\right)$$

$$= \sum_{a\in\mathcal{A}} \mathrm{Var}\left(Z^q_a\right)$$

$$= \sum_{a\in\mathcal{A}}\left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right] \tag{40}$$

$$= \left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right]p.$$

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

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

$$\left.\frac{p}{q^2\left(\frac{2p}{(q+2)(q+1)}\right)^{2\left(1-\frac{1}{q}\right)}}\left[\frac{1}{(q+1)(2q+1)} - \left(\frac{2}{(q+2)(q+1)}\right)^2\right]\right). \tag{41}$$

2.2 Manhattan (L_1)

With our general formulas for the asymptotic mean and variance given by Eqs. 31 and 41 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 metric $D^{(q)}_{ij}$. We demonstrate this with the example of the Manhattan metric for standard normal and standard uniform data, which is given by Eq. 1 by setting $q = 1$ (L_1).

2.2.1 Standard normal data

Using the mean given by Eq. 31 and substituting $q = 1$, we have the following for expected L_1 distance between two independently sample instances i and j in standard normal data

$$\mathrm{E}\left(D^{(1)}_{ij}\right) = \left(2\frac{\Gamma\left(\frac{1+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/1}$$

$$= \frac{2p}{\sqrt{\pi}}\Gamma(1) \tag{42}$$

$$= \frac{2p}{\sqrt{\pi}}.$$

We see in Eq. 42 that $D^{(1)}_{ij} \sim p$ 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 $D^{(1)}_{ij}$ given in

Eq. 31 leads to the following

$$\text{Var}\left(D_{ij}^{(1)}\right) = \frac{4^1 p}{1^2 \left(\frac{2^1 \Gamma\left(\frac{1}{2}(1)+\frac{1}{2}\right)}{\sqrt{\pi}} p\right)^{2\left(1-\frac{1}{1}\right)}} \left[\frac{\Gamma\left(1+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}(1)+\frac{1}{2}\right)}{\pi}\right]$$

$$= \frac{4p}{1}\left[\frac{\frac{1}{2}\Gamma\left(\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2(1)}{\pi}\right] \tag{43}$$

$$= 4p\left[\frac{1}{2} - \frac{1}{\pi}\right]$$

$$= \frac{2(\pi-2)p}{\pi}.$$

Similar to the mean given by Eq. 42, the limiting variance of $D_{ij}^{(1)}$ given by Eq. 43 grows on the order of feature dimension p, which implies that points become more dispersed as the dimension increases.

2.2.2   Standard uniform data

Using the mean given by Eq. 41 and substituting $q = 1$, we have the following for the expected L_1 distance between two independently sampled instances i and j in standard uniform data

$$\text{E}\left(D_{ij}^{(1)}\right) = \left(\frac{2p}{(1+2)(1+1)}\right)^{1/1}$$

$$= \frac{2p}{6} \tag{44}$$

$$= \frac{p}{3}.$$

Once again, we see that the mean of $D_{ij}^{(1)}$ given by Eq. 44 grows on the order of p just as in the case of standard normal data.

Substituting $q = 1$ into the formula given by Eq. 41 of the asymptotic variance of $D_{ij}^{(1)}$ leads to the following

$$\text{Var}\left(D_{ij}^{(1)}\right) = \frac{p}{1^2 \left(\frac{2p}{(1+2)(1+1)}\right)^{2\left(1-\frac{1}{1}\right)}} \left[\frac{1}{(1+1)(2(1)+1)} - \left(\frac{2}{(1+2)(1+1)}\right)^2\right]$$

$$= p\left[\frac{1}{6} - \frac{1}{9}\right] \tag{45}$$

$$= \frac{p}{18}.$$

As in the case of the L_1 metric on standard normal data, we have a variance given by Eq. 45 that grows on the order of p. The distances between points in high-dimensional uniform data become more widely dispersed with this metric.

2.3   Euclidean (L_2)

Analogous to the previous section, we demonstrate the usage of Eqs. 31 and 41 for the Euclidean metric for standard normal and standard uniform data, which is given by Eq. 1 by setting $q = 2$ (L_2).

2.3.1 Standard normal data

Using the mean given by Eq. 31 and substituting $q = 2$, we have the following for
expected L_2 distance between two independently sampled instances i and j in standard
normal data

$$
\begin{aligned}
\mathrm{E}\left(D_{ij}^{(2)}\right) &= \left(2\frac{\Gamma\left(\frac{2+1}{2}\right)}{\sqrt{\pi}}p\right)^{1/2} \\
&= \left(\frac{2p}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right)\right)^{1/2} \\
&= \sqrt{2p}.
\end{aligned}
\tag{46}
$$

In the case of L_2 on standard normal data, we see that the mean of $D_{ij}^{(2)}$ given by
Eq. 46 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 $D_{ij}^{(2)}$ given in
Eq. 31 leads to the following

$$
\begin{aligned}
\mathrm{Var}\left(D_{ij}^{(2)}\right) &= \frac{4^2 p}{2^2\left(\frac{2^2\Gamma\left(\frac{1}{2}(2)+\frac{1}{2}\right)}{\sqrt{\pi}}p\right)^{2\left(1-\frac{1}{2}\right)}}\left[\frac{\Gamma\left(2+\frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}(2)+\frac{1}{2}\right)}{\pi}\right] \\
&= \frac{16p}{4\left(\frac{4\Gamma\left(\frac{3}{2}\right)}{\sqrt{\pi}}p\right)}\left[\frac{\Gamma\left(\frac{5}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{3}{2}\right)}{\pi}\right] \\
&= 2\left[\frac{3}{4} - \frac{1}{4}\right] \\
&= 1.
\end{aligned}
\tag{47}
$$

Surprisingly, the asymptotic variance given by Eq. 47 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, an improved estimate of
the mean is given by Eq. 6. The lower dimensional estimate of the mean is as follows

$$
\begin{aligned}
\mathrm{E}\left(D_{ij}^{(2)}\right) &= \left(2\frac{\Gamma\left(\frac{2+1}{2}\right)}{\sqrt{\pi}}p - 1\right)^{1/2} \\
&= \left(\frac{2p}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) - 1\right)^{1/2} \\
&= \sqrt{2p - 1}.
\end{aligned}
\tag{48}
$$

For high dimensional data sets like gene expression [12, 13], which typically contain
thousands of genes (or features), it is clear that the magnitude of p will be sufficient to
use Eq. 46 since $\sqrt{2p} \approx \sqrt{2p - 1}$ in that case.

2.3.2 Standard uniform data

Using the mean given by Eq. 41 and substituting $q = 2$, we have the following for
expected L_2 distance between two independently sampled instances i and j in standard

uniform data

$$\mathrm{E}\left(D_{ij}^{(2)}\right) = \left(\frac{2p}{(2+2)(2+1)}\right)^{1/2}$$

$$= \left(\frac{2p}{12}\right)^{1/2} \tag{49}$$

$$= \sqrt{\frac{p}{6}}.$$

As in the case of standard normal data, the expected value of $D_{ij}^{(2)}$ given by Eq. 49 grows on the order of $\sqrt{p}$.

Substituting $q = 2$ into the formula for the asymptotic variance of $D_{ij}^{(2)}$ given in Eq. 41 leads to the following

$$\mathrm{Var}\left(D_{ij}^{(2)}\right) = \frac{p}{2^2\left(\frac{2p}{(2+2)(2+1)}\right)^{2\left(1-\frac{1}{2}\right)}}\left[\frac{1}{(2+1)(2(2)+1)} - \left(\frac{2}{(2+2)(2+1)}\right)^2\right]$$

$$= \frac{3}{2}\left[\frac{1}{15} - \frac{1}{36}\right] \tag{50}$$

$$= \frac{7}{120}.$$

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

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

$$\mathrm{E}\left(D_{ij}^{(2)}\right) = \left(\frac{2p}{(2+2)(2+1)} - \frac{7}{120}\right)^{1/2}$$

$$= \left(\frac{2p}{12} - \frac{7}{120}\right)^{1/2} \tag{51}$$

$$= \sqrt{\frac{p}{6} - \frac{7}{120}}.$$

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, 14] for scoring features.

2.4   Distribution of max-min normalized L_q metric

For Relief-based methods [1, 3, 14], the standard numeric diff metric is given by

$$\mathrm{d}_{ij}^{\mathrm{num}}(a) = \mathrm{diff}(a, (i, j)) = \frac{|X_{ia} - X_{ja}|}{\max(a) - \min(a)}, \tag{52}$$

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

In order to determine moments of asymptotic distance distributions induced by Eq. 52, we must first derive the asymptotic extreme value distributions of the attribute maximum and minimum. Although the exact distribution of the maximum or minimum requires an assumption about the data distribution, the Fisher-Tippett-Gnedenko Theorem allows us to generally categorize the extreme value distribution for a collection of independent and identically distributed random variables into one of three distributional families. Before stating the theorem, we first need the following definition

Definition 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 is derived as follows

$$
\begin{aligned}
\mathrm{P}[X_a^{\max} \leq x] &= \mathrm{P}\left[\max_{k \in \mathcal{I}}\{X_{ka}\} \leq x\right] \\
&= \mathrm{P}[X_{1a} \leq x, X_{2a} \leq x, \ldots, X_{ma} \leq x] \\
&= \prod_{k=1}^{m} \mathrm{P}[X_{ka} \leq x] \\
&= \prod_{k=1}^{m} F_X(x) \\
&= [F_X(x)]^m.
\end{aligned}
\tag{53}
$$

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

$$F_{\max}(x) = [F_X(x)]^m. \tag{54}$$

Differentiating the distribution function given by Eq. 54 gives us the following density function for the distribution of the maximum

$$
\begin{aligned}
f_{\max}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{\max}(x) \\
&= \frac{\mathrm{d}}{\mathrm{d}x} [F_X(x)]^m \\
&= m[F_X(x)]^{m-1} f_X(x).
\end{aligned}
\tag{55}
$$

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

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

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

$$
F_{\min}(x) = 1 - [1 - F_X(x)]^m.
\tag{57}
$$

Differentiating the distribution function given by Eq. 57 gives us the following density function for the distribution of the minimum

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

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

$$
\begin{aligned}
\mu_{\max}^{(1)}(m) = \mathrm{E}(X_a^{\max}) &= \int_{-\infty}^{\infty} x f_{\max}(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x\left(m[F_X(x)]^{m-1} f_X(x)\right)\mathrm{d}x \\
&= m \int_{-\infty}^{\infty} x f_X(x)[F_X(x)]^{m-1}\mathrm{d}x.
\end{aligned}
\tag{59}
$$

$$
\begin{aligned}
\mu_{\max}^{(2)}(m) = \mathrm{E}[(X_a^{\max})^2] &= \int_{-\infty}^{\infty} x^2 f_{\max}(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x^2\left(m[F_X(x)]^{m-1} f_X(x)\right)\mathrm{d}x \\
&= m \int_{-\infty}^{\infty} x^2 f_X(x)[F_X(x)]^{m-1}\mathrm{d}x
\end{aligned}
\tag{60}
$$

$$
\sigma_{\max}^2(m) = \mu_{\max}^{(2)}(m) - \left[\mu_{\max}^{(1)}(m)\right]^2
\tag{61}
$$

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

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

$$\mu_{\min}^{(2)}(m) = \mathrm{E}[(X_a^{\min})^2] = \int_{-\infty}^{\infty} x^2 f_{\min}(x) \mathrm{d}x$$

$$= \int_{-\infty}^{\infty} x^2 \left(m[F_X(x)]^{m-1} f_X(x) \right) \mathrm{d}x \qquad (63)$$

$$= m \int_{-\infty}^{\infty} x^2 f_X(x)[F_X(x)]^{m-1} \mathrm{d}x,$$

and

$$\sigma_{\min}^2(m) = \mu_{\min}^{(2)}(m) - \left[\mu_{\min}^{(1)}(m)\right]^2. \qquad (64)$$

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

$$\mathrm{E}(X_a^{\max} - X_a^{\min}) = \mathrm{E}(X_a^{\max}) - \mathrm{E}(X_a^{\min})$$
$$= \mu_{\max}^{(1)}(m) - \mu_{\min}^{(1)}(m). \qquad (65)$$

For a data distribution that has zero skewness and has support that is symmetric about 0, the result given by Eq. 65 can be simplified to the following expression

$$\mathrm{E}(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m). \qquad (66)$$

For large samples ($m \gg 1$), the covariance between the sample maximum and minimum is approximately zero [15]. Therefore, the variance of the attribute range of a sample of size m is given by the following

$$\mathrm{Var}(X_a^{\max} - X_a^{\min}) \approx \mathrm{Var}(X_a^{\max}) + \mathrm{Var}(X_a^{\min})$$
$$= \sigma_{\max}^2(m) + \sigma_{\min}^2(m). \qquad (67)$$

Under the assumption of zero skewness and support that is symmetric about 0, the result given by Eq. 67 becomes the following

$$\mathrm{Var}(X_a^{\max} - X_a^{\min}) = 2\mathrm{Var}(X_a^{\max})$$
$$= 2\sigma_{\max}^2. \qquad (68)$$

Let $\mu_{D_{ij}^{(q)}}$ and $\sigma_{D_{ij}^{(q)}}^2$ denote the mean and variance given by Eq. 18. Furthermore, let $D_{ij}^{(q*)}$ denote the max-min normalized distance between instances i and j that is induced by the metric given by Eq. 52. Then the mean of the max-min normalized distance distribution is given by the following

$$\mu_{D_{ij}^{(q*)}} = \mathrm{E}\left[\left(\sum_{a \in \mathcal{A}} \left(\frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{1/q} \right]$$

$$\approx \frac{1}{\mathrm{E}(X_a^{\max} - X_a^{\min})} \mathrm{E}\left[\left(\sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \right)^{1/q} \right] \qquad (69)$$

$$= \frac{\mu_{D_{ij}^{(q)}}}{\mathrm{E}(X_a^{\max}) - \mathrm{E}(X_a^{\min})}$$

$$= \frac{\mu_{D_{ij}^{(q)}}}{\mu_{\max}^{(1)} - \mu_{\min}^{(1)}}.$$

The variance of the max-min normalized distance distribution is given by the following 337

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

(70)

With the results given by Eqs. 69 and 70, we have the following generalized estimate 338
for the asymptotic distribution of the max-min normalized distance distribution 339

$$
D_{ij}^{(q*)} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{\mu_{\max}^{(1)}(m)-\mu_{\min}^{(1)}(m)}, \frac{\sigma^2_{D_{ij}^{(q)}}}{\mu_{\max}^{(2)}(m)-2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m)+\mu_{\min}^{(2)}(m)}\right). \quad (71)
$$

For data with zero skewness and support that is symmetric about 0, the expected 340
sample maximum is the additive inverse of the expected sample minimum. This allows 341
us to express the formula given by Eq. 69 exclusively in terms of the expected maximum. 342
This result is given by the following 343

$$
\mu_{D_{ij}^{(q*)}} \approx \frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}. \tag{72}
$$

A similar substitution gives us the following expression for the variance of the max-min 344
normalized distance distribution 345

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

(73)

Therefore, the asymptotic distribution of the max-min normalized distance distribu- 346

tion is given by the following

$$D_{ij}^{(q*)} \dot\sim \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}, \frac{\sigma_{D_{ij}^{(q)}}^2}{2\left(\sigma_{\max}^2(m) + \left[\mu_{\max}^{(1)}(m)\right]^2\right)}\right). \tag{74}$$

2.4.1 Standard normal data

Standard normal data has zero skewness and has support that is symmetric about 0.
This implies that the mean and variance of the distribution of sample range can be
expressed exclusively in terms of the sample maximum. Given the nature of the density
function of the sample maximum for sample size m, the integration required to determine
the moments given by Eqs. 59 and 60 is not possible. These moments can either be
approximated numerically or we can use extreme value theory to determine the form of
the asymptotic distribution of the sample maximum. Using the latter method, we will
show that the asymptotic distribution of the sample maximum for standard normal data
is in the Gumbel family. Let $c_m = -\Phi^{-1}\left(\frac{1}{m}\right)$ and $b_m = \frac{1}{c_m}$. Using Taylor's Theorem,
we have the following expansion

$$\begin{aligned}
\log\Phi(-c_m - b_m x) &= \log\Phi(-c_m) - b_m x \frac{\phi(-c_m)}{\Phi(-c_m)} + \mathcal{O}(b_m^2 x^2) \\
&= \log\left(\frac{1}{m}\right) - x\frac{\phi(-c_m)}{c_m\Phi(-c_m)} + \mathcal{O}(b_m^2 x^2).
\end{aligned} \tag{75}$$

In order to simplify the right-hand side of Eq. 75, we will use the well known Mills
Ratio Bounds [16] given by the following

$$1 \le \frac{\phi(x)}{x\Phi(-x)} \le 1 + \frac{1}{x^2} \quad, x > 0. \tag{76}$$

The inequalities given by Eq. 76 show that $\frac{\phi(x)}{x\Phi(-x)} \to 1$ as $x \to \infty$. This implies that
$\frac{\phi(c_m)}{c_m\Phi(-c_m)} \to 1$ as $m \to \infty$ since $c_m = -\Phi^{-1}\left(\frac{1}{m}\right) \to \infty$ as $m \to \infty$. This gives us the
following approximation of the right-hand side of Eq. 75

$$\begin{aligned}
\log\Phi(-c_m - b_m x) &\approx \log\left(\frac{1}{m}\right) - x + \mathcal{O}(b_m^2 x^2) \\
\Rightarrow \Phi(-c_m - b_m x) &\approx \frac{1}{m}e^{-x+\mathcal{O}(b_m^2 x^2)} \\
\Rightarrow \Phi(c_m + b_m x) &\approx 1 - \frac{1}{m}e^{-x+\mathcal{O}(b_m^2 x^2)}.
\end{aligned} \tag{77}$$

Using the result given by Eq. 77, we have the following

$$P\left(\frac{X_a^{\max} - c_m}{b_m} \leq x\right) = P(X_a^{\max} \leq c_m + b_m x)$$

$$= \Phi^m(c_m + b_m x)$$

$$\approx \left(1 - \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)}\right)^m$$

$$= \left(1 - \frac{1}{m} e^{-x + \mathcal{O}\left(\frac{1}{c_m^2} x^2\right)}\right)^m \qquad (78)$$

$$\approx \left(1 - \frac{1}{m} e^{-x}\right)^m$$

$$\Rightarrow \lim_{m \to \infty} P\left(\frac{X_a^{\max} - c_m}{b_m} \leq x\right) = \lim_{m \to \infty} \left(1 - \frac{1}{m} e^{-x}\right)^m$$

$$= e^{-e^{-x}}.$$

The right-hand side of Eq. 78 is the cumulative distribution function of the standard Gumbel distribution. The mean of the asymptotic distribution is given by the following

$$E(X_a^{\max}) = \mu_{\max}^{(1)} = -\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)}. \qquad (79)$$

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

$$\widetilde{\mu}_{\max} = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right). \qquad (80)$$

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

$$\mathrm{Var}(X_a^{\max}) = \frac{\pi^2}{6}\left(\frac{1}{-\Phi^{-1}\left(\frac{1}{m}\right)}\right)^2. \qquad (81)$$

For typical sample sizes m in high-dimensional spaces, the variance estimate given by Eq. 81 exceeds the variance of the sample maximum significantly. Using the fact that $-\Phi^{-1}\left(\frac{1}{m}\right) \sim \sqrt{2\log(m)}$ [17] and $\frac{1}{2\log(m)} \leq \left(\frac{1}{-\Phi^{-1}\left(\frac{1}{m}\right)}\right)^2$ for $m \geq 2$, we can get a more accurate approximation of the variance with the following

$$\sigma_{\max}^2(m) = \mathrm{Var}(X_a^{\max}) \approx \frac{\pi^2}{6}\left(\frac{1}{\sqrt{2\log(m)}}\right)^2$$

$$= \frac{\pi^2}{12\log(m)}. \qquad (82)$$

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

$$E(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m) = 2\left[-\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)}\right]. \qquad (83)$$

It is well known that the sample extremes from the standard normal distribution are approximately uncorrelated for large sample size m [15]. This implies that we can

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

$$
\begin{aligned}
\text{Var}(X_a^{\max} - X_a^{\min}) &\approx \text{Var}(X_a^{\max}) + \text{Var}(X_a^{\min}) \\
&= \sigma_{\max}^2(m) + \sigma_{\min}^2(m) \\
&= 2\sigma_{\max}^2(m) \\
&\approx 2\left(\frac{\pi^2}{2\log(m)}\right) \\
&= \frac{\pi^2}{6\log(m)}.
\end{aligned}
\tag{84}
$$

For the purpose of approximating the mean and variance of the max-min normalized distance distribution, the formula for the median of the distribution of the attribute maximum yields more accurate results. That is, the approximation of the expected maximum given by Eq. 79 overestimates the sample maximum. The formula for the median of the sample maximum, given by Eq. 80, provides a more accurate estimate of this sample extreme. Therefore, the following estimate for the mean of the attribute range will be used instead

$$
\text{E}(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m) \approx 2\left[\frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right)\right].
\tag{85}
$$

We have already determined that $\mu_{D_{ij}^{(q)}}$ and $\sigma_{D_{ij}^{(q)}}^2$ are given by Eq. 31. Using the results given by Eqs. 85 and 84 and the general formulas for the mean and variance of the max-min normalized distance distribution given in Eq. 74, this leads us to the following asymptotic estimate for the distribution of the max-min normalized distances for standard normal data

$$
D_{ij}^{(q*)} \overset{\cdot}{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}, \frac{6\log(m)\sigma_{D_{ij}^{(q)}}^2}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m)\right]^2\log(m)}\right).
\tag{86}
$$

2.4.2   Standard uniform data

Standard uniform data does not have support that is symmetric about 0. Due to the simplicity of the density function, however, we can derive the distribution of the maximum and minimum of a sample of size m explicitly. Using the general forms of the distribution functions of the maximum and minimum given by Eqs. 54 and 57, we have the following distribution functions for standard uniform data

$$
F_{\max}(x) = x^m
\tag{87}
$$

and

$$
F_{\min}(x) = 1 - (1 - x)^m.
\tag{88}
$$

Using the general forms of the density functions of the maximum and minimum given by Eqs. 55 and 58, we have the following density functions for standard uniform data

$$
f_{\max}(x) = mx^{m-1}
\tag{89}
$$

and

$$
f_{\min}(x) = m(1 - x)^{m-1}
\tag{90}
$$

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

$$
\begin{aligned}
\mathrm{E}(X_a^{\mathrm{max}}) = \mu_{\mathrm{max}}^{(1)}(m) &= \int_0^1 x f_{\mathrm{max}}(x)\mathrm{d}x \\
&= \int_0^1 x[mx^{m-1}]\mathrm{d}x \\
&= \frac{m}{m+1}
\end{aligned} \tag{91}
$$

and

$$
\begin{aligned}
\mathrm{E}(X_a^{\mathrm{min}}) = \mu_{\mathrm{min}}^{(1)}(m) &= \int_0^1 x f_{\mathrm{min}}(x)\mathrm{d}x \\
&= \int_0^1 x[m(1-x)^{m-1}]\mathrm{d}x \\
&= \frac{1}{m+1}.
\end{aligned} \tag{92}
$$

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

$$
\begin{aligned}
\mathrm{E}[(X_a^{\mathrm{max}} - X_a^{\mathrm{min}})^2] &= \mathrm{E}[(X_a^{\mathrm{max}})^2 - 2X_a^{\mathrm{max}}X_a^{\mathrm{min}} + (X_a^{\mathrm{min}})^2] \\
&= \mathrm{E}[(X_a^{\mathrm{max}})^2] - 2\mathrm{E}(X_a^{\mathrm{max}})\mathrm{E}(X_a^{\mathrm{min}}) + \mathrm{E}[(X_a^{\mathrm{min}})^2] \\
&= \mu_{\mathrm{max}}^{(2)}(m) - 2\mu_{\mathrm{max}}^{(1)}(m)\mu_{\mathrm{min}}^{(1)}(m) + \mu_{\mathrm{min}}^{(2)}(m) \\
&= \int_0^1 x^2[mx^{m-1}]\mathrm{d}x - 2\left(\frac{m}{m+1}\right)\left(\frac{1}{m+1}\right) \\
&\qquad + \int_0^1 x^2[m(1-x)^{m-1}]\mathrm{d}x \\
&= \frac{m}{m+2} - \frac{2m}{(m+1)^2} + \frac{2}{(m+1)(m+2)} \\
&= \frac{m^3 - m + 2}{(m+2)(m+1)^2}.
\end{aligned} \tag{93}
$$

Using the general formulas given in Eq. 71 and the mean $(\mu_{D_{ij}^{(q)}})$ and variance $(\sigma^2_{D_{ij}^{(q)}})$ given by Eq. 41, we have the following asymptotic estimate for the max-min normalized distance distribution for standard uniform data

$$
D_{ij}^{(q*)} \overset{.}{\sim} \mathcal{N}\left(\frac{(m+1)\mu_{D_{ij}^{(q)}}}{m-1}, \frac{(m+2)(m+1)^2\sigma^2_{D_{ij}^{(q)}}}{m^3 - m + 2}\right). \tag{94}
$$

2.5   Normalized Manhattan $(q = 1)$

Using the general asymptotic results for mean and variance given by Eqs. 86 and 94 for any value of $q \in \mathbb{N}$, we can substitute a particular value of q in order to determine a more specified asymptotic distance distribution for $D_{ij}^{(q*)}$. The following results are for the max-min normalized Manhattan $(q = 1)$ metric for both standard normal and standard uniform data.

2.5.1 Standard normal data

Substituting $q = 1$ into Eq. 86, we have the following for standard normal data

$$
\begin{aligned}
\mathrm{E}\left(D_{ij}^{(1*)}\right) &= \frac{\mu_{D_{ij}^{(1)}}}{2\mu_{\max}^{(1)}(m)} \\
&= \frac{p}{\sqrt{\pi}\mu_{\max}^{(1)}(m)},
\end{aligned}
\tag{95}
$$

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 80.

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

$$
\begin{aligned}
\mathrm{Var}\left(D_{ij}^{(1*)}\right) &= \frac{6\log(m)\sigma_{D_{ij}^{(1)}}^2}{\pi^2 + 24\left[\mu_{\max}^{(1)}\right]^2\log(m)} \\
&= \frac{12p(\pi-2)\log(m)}{\pi\left(\pi^2 + 24\left[\mu_{\max}^{(1)}\right]^2\log(m)\right)},
\end{aligned}
\tag{96}
$$

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 80.

2.5.2 Standard uniform data

Substituting $q = 1$ into Eq. 94, we have the following for standard uniform data

$$
\begin{aligned}
\mathrm{E}\left(D_{ij}^{(1*)}\right) &= \frac{(m+1)\mu_{D_{ij}^{(1)}}}{m-1} \\
&= \frac{(m+1)p}{3(m-1)}.
\end{aligned}
\tag{97}
$$

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

$$
\begin{aligned}
\mathrm{Var}\left(D_{ij}^{(1*)}\right) &= \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(1)}}^2}{m^3 - m + 2} \\
&= \frac{(m+2)(m+1)^2 p}{18(m^3 - m + 2)}.
\end{aligned}
\tag{98}
$$

2.6 Normalized Euclidean ($q = 2$)

Analogous to the previous section, we demonstrate the usage of Eqs. 86 and 94 for the
max-min normalized Euclidean ($q = 2$) metric for both standard normal and standard
uniform data.

2.6.1 Standard normal data

Substituting $q = 2$ into Eq. 86, we have the following for standard normal data

$$
\begin{aligned}
\mathrm{E}\left(D_{ij}^{(2*)}\right) &= \frac{\mu_{D_{ij}^{(2)}}}{2\mu_{\max}^{(1)}(m)} \\
&= \frac{\sqrt{2p-1}}{2\mu_{\max}^{(1)}(m)},
\end{aligned}
\tag{99}
$$

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 80.

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

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

$$= \frac{6\log(m)}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m)\right]^2 \log(m)},$$

(100)

where $\mu_{\max}^{(1)}(m)$ is given by Eq. 80.

2.6.2 Standard uniform data

Substituting $q = 2$ into Eq. 94, we have the following for standard uniform data

$$\text{E}\left(D_{ij}^{(2*)}\right) = \frac{(m+1)\mu_{D_{ij}^{(2)}}}{m-1}$$

$$= \sqrt{\frac{p}{6} - \frac{7}{120}}\left(\frac{m+1}{m-1}\right).$$

(101)

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

$$\text{Var}\left(D_{ij}^{(2*)}\right) = \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(2)}}^2}{m^3 - m + 2}$$

$$= \frac{7(m+2)(m+1)^2}{120(m^3 - m + 2)}.$$

(102)

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

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

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

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

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

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

3 GWAS distance distributions

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

$$X_{ia} = \begin{cases} 0 & \text{if there are no minor alleles at locus } a, \\ 1 & \text{if there is 1 minor allele at locus } a, \\ 2 & \text{if there are 2 minor alleles at locus } a. \end{cases} \quad (103)$$

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

$$d_{ij}^{\mathrm{GM}}(a) = \begin{cases} 0 & \text{if } X_{ia} \neq X_{ja}, \\ 1 & \text{otherwise} \end{cases} \quad (104)$$

and

$$d_{ij}^{\mathrm{AM}}(a) = \frac{1}{2}\big|X_{ia} - X_{ja}\big|. \quad (105)$$

A more informative metric must take into account whether differences in allele frequency at a particular locus a result in transitions or transversions. A metric that accounts for transitions (Ti) and transversions (Tv) was introduced in [9]. This metric is given by the following

$$d_{ij}^{\mathrm{TiTv}}(a) = \begin{cases} 0 & \text{if } X_{ia} = X_{ja} \text{ and Ti/Tv}, \\ 1/4 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Ti}, \\ 1/2 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Tv}, \\ 3/4 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Ti}, \\ 1 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Tv}. \end{cases} \quad (106)$$

With any of the three metrics given by Eqs. 104 - 106, we compute the pairwise distance between two instances i and j using Eq. 1 with $q = 1$. Assuming that all data entries X_{ia} are independent and identically distributed, we have already shown that the distribution of pairwise distances is asymptotically normal regardless of data distribution and value of q. Therefore, the distance distributions induced by each of the GWAS metrics given by Eqs. 104 - 106 are asymptotically normal. Thus, 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 expected value of the GM metric is given by the following

$$
\begin{aligned}
\mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right] &= \sum_{k=0}^{1} k \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = k\right] \\
&= 0 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 0\right] + 1 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 1\right] \\
&= \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = 1\right] \\
&= 2\mathrm{P}[X_{ia} = 0, X_{ja} = 1] + 2\mathrm{P}[X_{ia} = 1, X_{ja} = 2] + 2\mathrm{P}[X_{ia} = 0, X_{ja} = 2] \\
&= 4(1 - f_a)^3 f_a + 4(1 - f_a)f_a^3 + 2(1 - f_a)^2 f_a^2 \\
&= 2\left[2(1 - f_a)^3 f_a + 2(1 - f_a)f_a^3 + (1 - f_a)^2 f_a^2\right] \\
&= 2F(a),
\end{aligned}
\tag{107}
$$

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

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

$$
\begin{aligned}
\mathrm{E}\left(D_{ij}^{\mathrm{GM}}\right) &= \mathrm{E}\left(\sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\mathrm{GM}}(a)\right) \\
&= \sum_{a \in \mathcal{A}} \mathrm{E}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right] \\
&= 2\sum_{a \in \mathcal{A}} F(a).
\end{aligned}
\tag{108}
$$

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

$$
\begin{aligned}
\mathrm{E}\left[\left(D_{ij}^{\mathrm{GM}}\right)^2\right] &= \mathrm{E}\left[\left(\sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\mathrm{GM}}(a)\right)^2\right] \\
&= \mathrm{E}\left[\sum_{a \in \mathcal{A}}\left(\mathrm{d}_{ij}^{\mathrm{GM}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \mathrm{d}_{ij}^{\mathrm{GM}}(r) \cdot \mathrm{d}_{ij}^{\mathrm{GM}}(s)\right] \\
&= \sum_{a \in \mathcal{A}}\left(\sum_{k=0}^{1} k^2 \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(a) = k\right]\right) \\
&\quad + 2\sum_{a \in \mathcal{A}} \sum_{s \leq r-1}\left(\sum_{k=0}^{1} k \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(r) = k\right]\right) \cdot \left(\sum_{k=0}^{1} k \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{GM}}(s) = k\right]\right) \\
&= 2\sum_{a \in \mathcal{A}} F(a) + 8\sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r,s\}} F(\lambda),
\end{aligned}
\tag{109}
$$

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

Using the moments given by Eqs. 108 and 109, the variance is computed as follows

$$\text{Var}\left(D_{ij}^{\text{GM}}\right) = \text{E}\left[\left(D_{ij}^{\text{GM}}\right)^2\right] - \left[\text{E}\left(D_{ij}^{\text{GM}}\right)\right]^2$$

$$= 2\sum_{a\in\mathcal{A}} F(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F(\lambda) - 4\left(\sum_{a\in\mathcal{A}} F(a)\right)^2$$

$$= 2\sum_{a\in\mathcal{A}} F(a) - 4\sum_{a\in\mathcal{A}} F^2(a)$$

$$= 2\sum_{a\in\mathcal{A}} F(a)[1 - 2F(a)],$$

(110)

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

With the mean and variance estimates given by Eqs. 108 and 110, the asymptotic GM distance distribution is given by the following

$$D_{ij}^{\text{GM}} \overset{\cdot}{\sim} \mathcal{N}\left(2\sum_{a\in\mathcal{A}} F(a), 2\sum_{a\in\mathcal{A}} F(a)[1 - 2F(a)]\right),$$

(111)

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

3.2   AM distance distribution

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

$$\text{E}\left[\text{d}_{ij}^{\text{AM}}(a)\right] = \sum_{k\in\mathcal{D}} k \cdot \text{P}\left[\text{d}_{ij}^{\text{AM}}(a) = k\right]$$

$$= 0 \cdot \text{P}\left[\text{d}_{ij}^{\text{AM}}(a) = 0\right] + \frac{1}{2} \cdot \text{P}\left[\text{d}_{ij}^{\text{AM}}(a) = \frac{1}{2}\right] + 1 \cdot \text{P}\left[\text{d}_{ij}^{\text{AM}}(a) = 1\right]$$

$$= \frac{1}{2}\left(2\text{P}\left[X_{ia} = 0, X_{ja} = 1\right] + 2\text{P}\left[X_{ia} = 1, X_{ja} = 2\right]\right)$$

$$+ 2\text{P}\left[X_{ia} = 0, X_{ja} = 2\right]$$

$$= \text{P}\left[X_{ia} = 0, X_{ja} = 1\right] + \text{P}\left[X_{ia} = 1, X_{ja} = 2\right] + 2\text{P}\left[X_{ia} = 0, X_{ja} = 2\right]$$

$$= 2(1-f_a)^3 f_a + 2(1-f_a)f_a^3 + 2(1-f_a)^2 f_a^2$$

$$= 2\left[(1-f_a)^3 f_a + (1-f_a)f_a^3 + (1-f_a)^2 f_a^2\right]$$

$$= 2F(a),$$

(112)

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

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

$$\text{E}\left(D_{ij}^{\text{AM}}\right) = \text{E}\left(\sum_{a\in\mathcal{A}} \text{d}_{ij}^{\text{AM}}(a)\right)$$

$$= \sum_{a\in\mathcal{A}} \text{E}\left[\text{d}_{ij}^{\text{AM}}(a)\right]$$

(113)

$$= 2\sum_{a\in\mathcal{A}} F(a).$$

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

$$\mathrm{E}\left[\left(D_{ij}^{\mathrm{AM}}\right)^2\right] = \mathrm{E}\left[\left(\sum_{a\in\mathcal{A}} \mathrm{d}_{ij}^{\mathrm{AM}}(a)\right)^2\right]$$

$$= \mathrm{E}\left[\sum_{a\in\mathcal{A}}\left(\mathrm{d}_{ij}^{\mathrm{AM}}(a)\right)^2\right] + 2\mathrm{E}\left[\sum_{r\in\mathcal{A}}\sum_{s\leq r-1} \mathrm{d}_{ij}^{\mathrm{AM}}(r)\cdot \mathrm{d}_{ij}^{\mathrm{AM}}(s)\right]$$

$$= \sum_{a\in\mathcal{A}}\left(\sum_{k\in\mathcal{D}} k^2\cdot\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(a)=k\right]\right)$$

$$+ 2\sum_{a\in\mathcal{A}}\sum_{s\leq r-1}\left(\sum_{k\in\mathcal{D}} k\cdot\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(r)=k\right]\right)\cdot\left(\sum_{k\in\mathcal{D}} k\cdot\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{AM}}(s)=k\right]\right)$$

$$= \sum_{a\in\mathcal{A}} G(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F(\lambda),$$

$$(114)$$

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

Using the moments given by Eqs. 113 and 114, the variance is computed as follows

$$\mathrm{Var}\left(D_{ij}^{\mathrm{AM}}\right) = \mathrm{E}\left[\left(D_{ij}^{\mathrm{AM}}\right)^2\right] - \left[\mathrm{E}\left(D_{ij}^{\mathrm{AM}}\right)\right]^2$$

$$= \sum_{a\in\mathcal{A}} G(a) + 8\sum_{r\in\mathcal{A}}\sum_{s\leq r-1}\prod_{\lambda\in\{r,s\}} F(\lambda) - 4\left(\sum_{a\in\mathcal{A}} F(a)\right)^2$$

$$= \sum_{a\in\mathcal{A}} G(a) - 4\sum_{a\in\mathcal{A}} F^2(a)$$

$$= \sum_{a\in\mathcal{A}} \left[G(a) - 4F^2(a)\right],$$

$$(115)$$

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

With the mean and variance estimates given by Eqs. 113 and 115, the asymptotic
AM distance distribution is given by the following

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

$$(116)$$

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

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 an j is that we know whether the nucleotides
associated with a particular locus a are both purines (PuPu), purine and pyrimidine
(PuPy), or both pyrimidines (PyPy). A diagram showing possible transitions and
transversions that may occur is given by Fig. 3. Purines (A and G) and pyrimidines (C
and T) are shown at the top and bottom, respectively. Transitions occur in the cases of
PuPu and PyPy, while transversion occur only with PuPy encoding.
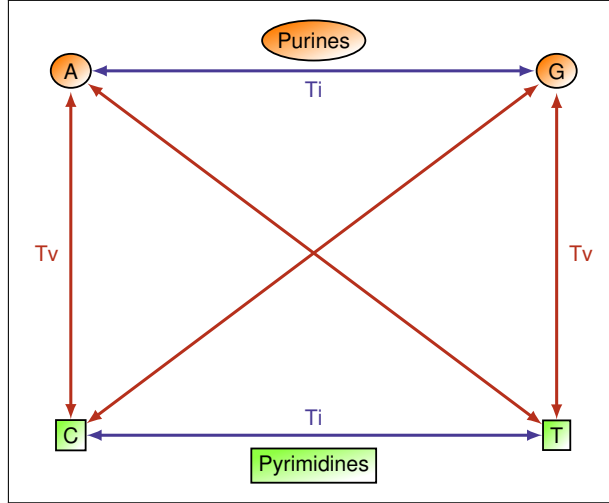
Fig 3. Purines (A and G) and pyrimidines (C and T) are shown. Transitions occur when a mutation involves purine-to-purine or pyrimidine-to-pyrimidine insertion. Transversions occur when a purine-to-pyrimidine or pyrimidine-to-purine insertion happens, which is a more extreme case. There are visibly more possibilities for transversions to occur than there are transitions, but there are about twice as many transitions in real data.

This information is always given in a particular data set. Let γ_0, γ_1, and γ_2 denote the probabilities of PuPu, PuPy, and PyPy, respectively, for the p loci of data matrix X. In real data, there are approximately twice as many transitions as there are transversions. That is, the probability of a transition P(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

$$\gamma_0 + \gamma_1 + \gamma_2 = 1, \tag{117}$$

$$P(\text{Ti}) - \eta P(\text{Tv}) = 0. \tag{118}$$

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

$$P(\text{Tv}) = \gamma_1. \tag{119}$$

Using the constraints given by Eqs. 117 and 118, the probability of a transition occuring at locus a is computed as follows

$$P(\text{Ti}) = \gamma_0 + \gamma_2. \tag{120}$$

Also based on the constraints given by Eqs. 117 and 118, it is clear that we have $P(\text{Tv}) = \frac{1}{\eta+1}$ and $P(\text{Ti}) = \frac{\eta}{\eta+1}$. Without loss of generality, we then sample

$$\gamma_0 \sim \mathcal{U}\left(\varepsilon, \frac{\eta}{\eta+1} - \varepsilon\right), \tag{121}$$

where ε is some small positive real number.

Then it immediately follows that we have

$$\gamma_2 = \frac{\eta}{\eta+1} - \gamma_0. \tag{122}$$

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

$$y_a = \begin{cases} 0 & \text{if locus } a \text{ is PuPu,} \\ 1 & \text{if locus } a \text{ is PuPy,} \\ 2 & \text{if locus } a \text{ is PyPy.} \end{cases} \tag{123}$$

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

$$\begin{aligned}
\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = 0\right] &= \mathrm{P}\left[y_a = 0, X_{ia} = X_{ja}\right] \\
&\quad + \mathrm{P}\left[y_a = 1, X_{ia} = X_{ja}\right] \\
&\quad + \mathrm{P}\left[y_a = 2, X_{ia} = X_{ja}\right] \\
&= \gamma_0 \left[(1 - f_a)^2 + 4 f_a(1 - f_a) + f_a^2\right] \\
&\quad + \gamma_1 \left[(1 - f_a)^2 + 4 f_a(1 - f_a) + f_a^2\right] \\
&\quad + \gamma_2 \left[(1 - f_a)^2 + 4 f_a(1 - f_a) + f_a^2\right] \\
&= (\gamma_0 + \gamma_1 + \gamma_2) \left[(1 - f_a)^2 + 4 f_a(1 - f_a) + f_a^2\right] \\
&= (1 - f_a)^2 + 4 f_a(1 - f_a) + f_a^2.
\end{aligned} \tag{124}$$

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

$$\begin{aligned}
\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = \frac{1}{4}\right] &= 2\mathrm{P}\left[y_a = 0, X_{ia} = 0, X_{ja} = 1\right] \\
&\quad + 2\mathrm{P}\left[y_a = 0, X_{ia} = 1, X_{ja} = 2\right] \\
&\quad + 2\mathrm{P}\left[y_a = 2, X_{ia} = 0, X_{ja} = 1\right] \\
&\quad + 2\mathrm{P}\left[y_a = 2, X_{ia} = 1, X_{ja} = 2\right] \\
&= 4\gamma_0(1 - f_a)^3 f_a + 4\gamma_0 f_a^3(1 - f_a) + 4\gamma_2(1 - f_a)^3 f_a \\
&\quad + 4\gamma_2 f_a^3(1 - f_a) \\
&= 4\gamma_0 \left[(1 - f_a)^3 f_a + f_a^3(1 - f_a)\right] \\
&\quad + 4\gamma_2 \left[(1 - f_a)^3 f_a + f_a^3(1 - f_a)\right] \\
&= 4(\gamma_0 + \gamma_2) \left[(1 - f_a)^3 f_a + f_a^3(1 - f_a)\right].
\end{aligned} \tag{125}$$

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

$$\begin{aligned}
\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = \frac{1}{2}\right] &= 2\mathrm{P}\left[y_a = 1, X_{ia} = 0, X_{ja} = 1\right] \\
&\quad + 2\mathrm{P}\left[y_a = 1, X_{ia} = 1, X_{ja} = 2\right] \\
&= 4\gamma_1(1 - f_a)^3 f_a + 4\gamma_1 f_a^3(1 - f_a) \\
&= 4\gamma_1 \left[(1 - f_a)^3 f_a + f_a^3(1 - f_a)\right].
\end{aligned} \tag{126}$$

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

$$\begin{aligned}
\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = \frac{3}{4}\right] &= 2\mathrm{P}\left[y_a = 0, X_{ia} = 0, X_{ja} = 2\right] \\
&\quad + 2\mathrm{P}\left[y_a = 2, X_{ia} = 0, X_{ja} = 2\right] \\
&= 2\gamma_0(1 - f_a)^2 f_a^2 + 2\gamma_2(1 - f_a)^2 f_a^2 \\
&= 2(\gamma_0 + \gamma_2)(1 - f_a)^2 f_a^2.
\end{aligned} \tag{127}$$

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

$$
\begin{aligned}
\mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = 1\right] &= 2\mathrm{P}\left[y_a = 1, X_{ia} = 0, X_{ja} = 2\right] \\
&= 2\gamma_1 (1 - f_a)^2 f_a^2.
\end{aligned} \tag{128}
$$

Using Eqs. 124 - 128, we compute the expected TiTv distance between instances i and j as follows

$$
\begin{aligned}
\mathrm{E}\left(D_{ij}^{\mathrm{TiTv}}\right) &= \sum_{a \in \mathcal{A}}\left(\sum_{k \in \mathcal{D}} k \cdot \mathrm{P}\left[\mathrm{d}_{ij}^{\mathrm{TiTv}}(a) = k\right]\right) \\
&= (\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}}\left[(1 - f_a)^3 f_a + f_a^3(1 - f_a)\right] \\
&\quad + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} (1 - f_a)^2 f_a^2 \\
&= (\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a),
\end{aligned} \tag{129}
$$

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

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

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

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

Using the moments given by Eqs. 129 and 130, the variance is computed as follows

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

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

With the mean and variance estimates given by Eqs. 129 and 131, the asymptotic TiTv distance distribution is given by the following

$$
\begin{aligned}
D_{ij}^{\mathrm{TiTv}} \overset{\cdot}{\sim} \mathcal{N}\bigg(&(\gamma_0 + \gamma_2 + 2\gamma_1)\sum_{a \in \mathcal{A}} F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right]\sum_{a \in \mathcal{A}} G(a), \\
&\left[\frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1\right] \sum_{a \in \mathcal{A}} F(a) + \left[\frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1\right] \sum_{a \in \mathcal{A}} G(a) \\
&- \sum_{a \in \mathcal{A}} \left([\gamma_0 + \gamma_2 + 2\gamma_1]F(a) + \left[\frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1\right]G(a)\right)^2 \bigg),
\end{aligned}
\tag{132}
$$

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

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

$$
\bar{f}_a = \frac{1}{2}(l + u).
\tag{133}
$$

The maximum TiTv distance occurs at $\bar{f}_a = 0.5$ for any fixed Ti/Tv ratio η (Eq. 118), which is the inflection point about which the minor allele changes at locus a (Fig. 4). 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 ($\bar{f}_a \to 1$). 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. 5A). 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 $\eta = \mathrm{Ti/Tv} = 0.5, 1, 1.5,$ and 2 (Fig. 5C). The TiTv distance decreases slightly with increased $\eta = \mathrm{Ti/Tv}$. 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.

We also compared theoretical and sample moments as a function of $\eta = \mathrm{Ti/Tv}$ and $\bar{f}_a$ for the TiTv distance metric (Fig. 5B and D). We fixed $\bar{f}_a$ and computed the theoretical and simulated moments as a function of η (Fig. 5B). Theoretical average TiTv distance, given by Eq. 129, and simulated TiTv average distance are approximately equal as

η increases. Theoretical standard deviation, given by Eq. 131, 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. 5D). In this case, there is approximate agreement
with simulated and theoretical moments as $\bar{f}_a$ increases.
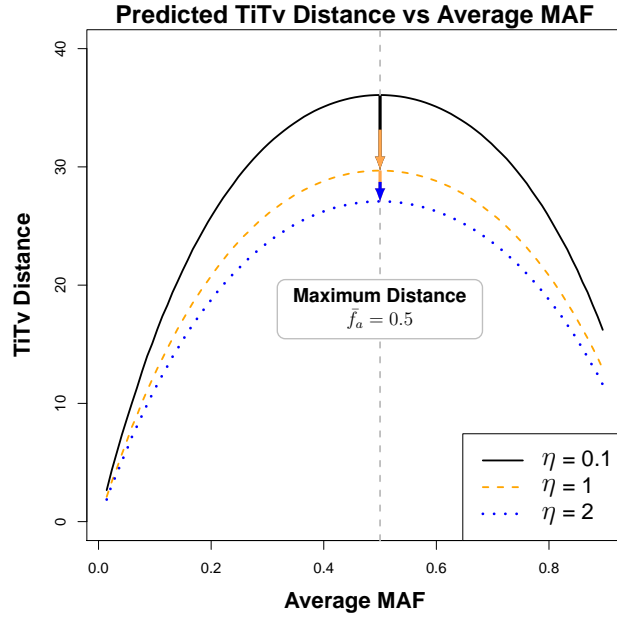
Fig 4. Predicted average TiTv distance as a function of average minor allele frequency $\bar{f}_a$ (see Eq. 133). Success probabilities f_a were drawn from a sliding window interval from 0.01 to 0.9 in increments of about 0.009. With $\eta = 0.1$, where η is the Ti/Tv ratio given by Eq. 117, Tv is ten 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.
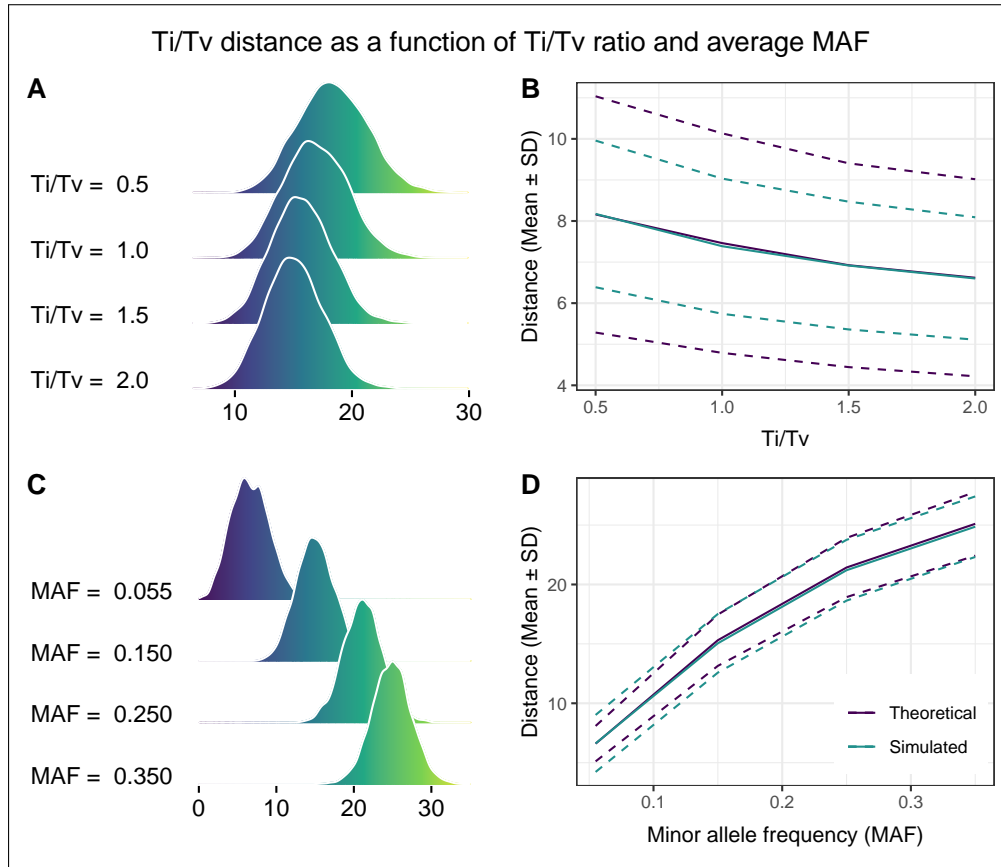
Fig 5. Density curves and moments of TiTv distance as a function of average MAF $\bar{f}_a$, given by Eq. 133, and Ti/Tv ratio η, given by Eq. 118. (**A**) For fixed $\eta = 2$, TiTv distance density is plotted as a function of increasing $\bar{f}_a = 0.055, 0.150, 0.250,$ and 0.350. TiTv distance increases as $\bar{f}_a$ approaches a maximum of 0.5, which means that there is about the same frequency of minor alleles as primary alleles at locus a. (**B**) Simulated and predicted mean $\pm$ SD are shown as a function of increasing Ti/Tv ratio η. Distance decreases as Tv becomes more frequent than Ti. Theoretical standard deviation is slightly larger than simulated, but the means are approximately the same. (**C**) For fixed $\bar{f}_a = 0.055$, TiTv distance density is plotted as a function of increasing $\eta = 0.5, 1, 1.5,$ and 2. TiTv distance decreases as η, the Ti/Tv ratio, increases. For $\eta = \mathrm{Ti/Tv} = 0.5$, there are twice as many transversions as there are transitions. On the other hand, $\eta = \mathrm{Ti/Tv} = 2$ indicates that there are half as many transversions as there are transitions. Since transversions encode a larger magnitude distance than transitions in Eq. 106, this behavior is expected. (**D**) Simulated and predicted mean $\pm$ SD are shown as a function of increasing average MAF $\bar{f}_a$. Distance increases as the number of minor alleles increases at each locus a. 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. 103)**	mean	$\boxed{2\sum_{a\in\mathcal{A}}F(a)}$ **(110)** where $F(a)=2(1-f_a)^3 f_a + 2f_a^3(1-f_a)+(1-f_a)^2 f_a^2$
	variance	$\boxed{2\sum_{a\in\mathcal{A}}F(a)[1-2F(a)]}$ **(110)** where $F(a)=2(1-f_a)^3 f_a + 2f_a^3(1-f_a)+(1-f_a)^2 f_a^2$
AM **(Eq. 104)**	mean	$\boxed{2\sum_{a\in\mathcal{A}}F(a)}$ **(115)** where $F(a)=(1-f_a)^3 f_a + f_a^3(1-f_a)+(1-f_a)^2 f_a^2$
	variance	$\boxed{\sum_{a\in\mathcal{A}}[G(a)-4F^2(a)]}$ **(115)** where $F(a)=2(1-f_a)^3 f_a + 2f_a^3(1-f_a)+(1-f_a)^2 f_a^2$ and $G(a)=(1-f_a)^3 f_a + f_a^3(1-f_a)+2(1-f_a)^2 f_a^2$
TiTv **(Eq. 105)**	mean	$\boxed{(\gamma_0+\gamma_2+2\gamma_1)\sum_{a\in\mathcal{A}}F(a)+\left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}}G(a)}$ **(131)** where $F(a)=(1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G(a)=(1-f_a)^2 f_a^2$
	mean	$\boxed{\begin{aligned}&\left[\frac{1}{4}(\gamma_0+\gamma_2)+\gamma_1\right]\sum_{a\in\mathcal{A}}F(a)+\left[\frac{9}{8}(\gamma_0+\gamma_2)+2\gamma_1\right]\sum_{a\in\mathcal{A}}G(a)\\&+\sum_{a\in\mathcal{A}}\left([\gamma_0+\gamma_2+2\gamma_1]F(a)+\left[\frac{3}{2}(\gamma_0+\gamma_2)+2\gamma_1\right]G(a)\right)^2\end{aligned}}$ **(131)** where $F(a)=(1-f_a)^3 f_a + f_a^3(1-f_a)$ and $G(a)=(1-f_a)^2 f_a^2$

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 [18]. These correlations are between different ROIs for a particular brain atlas [19]. Because the features are the ROIs themselves, this leads us to the following metric

$$\mathrm{d}_{ij}^{\mathrm{ROI}}(a) = \sum_{k\neq a}\left|A_{ka}^{(i)} - A_{ka}^{(j)}\right|. \tag{134}$$

where $A_{ka}^{(i)}$ and $A_{ka}^{(j)}$ are the correlations between ROI a and ROI k for instances i and j, respectively. In order for comparisons between different correlations to be possible, we first perform a Fisher r-to-z transform on the correlations. We then load all of the

transformed correlations into a $p(p-1) \times m$ matrix X (see Fig. 6). 562

$$
\text{ROI}_1 \left\{ \quad \text{ROI}_2 \left\{ \quad \vdots \quad \text{ROI}_p \left\{ \quad
\begin{array}{c}
\begin{array}{cccccc}
\boldsymbol{I_1} & \boldsymbol{I_2} & \boldsymbol{I_3} & \boldsymbol{I_4} & \cdots & \boldsymbol{I_m}
\end{array} \\
\begin{bmatrix}
\hat{A}_{12}^{(1)} & \hat{A}_{12}^{(2)} & \hat{A}_{12}^{(3)} & \hat{A}_{12}^{(4)} & \cdots & \hat{A}_{12}^{(m)} \\
\hat{A}_{13}^{(1)} & \hat{A}_{13}^{(2)} & \hat{A}_{13}^{(3)} & \hat{A}_{13}^{(4)} & \cdots & \hat{A}_{13}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{A}_{1p}^{(1)} & \hat{A}_{1p}^{(2)} & \hat{A}_{1p}^{(3)} & \hat{A}_{1p}^{(4)} & \cdots & \hat{A}_{1p}^{(m)} \\
\hat{A}_{21}^{(1)} & \hat{A}_{21}^{(2)} & \hat{A}_{21}^{(3)} & \hat{A}_{21}^{(4)} & \cdots & \hat{A}_{21}^{(m)} \\
\hat{A}_{23}^{(1)} & \hat{A}_{23}^{(2)} & \hat{A}_{23}^{(3)} & \hat{A}_{23}^{(4)} & \cdots & \hat{A}_{23}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{A}_{2p}^{(1)} & \hat{A}_{2p}^{(2)} & \hat{A}_{2p}^{(3)} & \hat{A}_{2p}^{(4)} & \cdots & \hat{A}_{2p}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{A}_{p1}^{(1)} & \hat{A}_{p1}^{(2)} & \hat{A}_{p1}^{(3)} & \hat{A}_{p1}^{(4)} & \cdots & \hat{A}_{p1}^{(m)} \\
\hat{A}_{p2}^{(1)} & \hat{A}_{p2}^{(2)} & \hat{A}_{p2}^{(3)} & \hat{A}_{p2}^{(4)} & \cdots & \hat{A}_{p2}^{(m)} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
\hat{A}_{p,p-1}^{(1)} & \hat{A}_{p,p-1}^{(2)} & \hat{A}_{p,p-1}^{(3)} & \hat{A}_{p,p-1}^{(4)} & \cdots & \hat{A}_{p,p-1}^{(m)}
\end{bmatrix}
\end{array}
\right. \right. \right. = \boldsymbol{X}
$$

Fig 6. 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. 563

We further transform the data matrix X by standardizing so that each of the m 564 columns has zero mean and unit variance. Therefore, the data in matrix X are standard 565 normal. Recall from Eqs. 42 and 43, that the mean and variance of the Manhattan 566 ($q = 1$) distance distribution for standard normal data are $\frac{2p}{\sqrt{\pi}}$ and $\frac{2(\pi-2)p}{\pi}$, respectively. 567 This allows us to easily derive the expected pairwise distance between instances i and j 568 in rs-fMRI data as follows 569

$$
\begin{aligned}
\mathrm{E}(D_{ij}^{\text{fMRI}}) &= \mathrm{E}\left(\sum_{a \in \mathcal{A}} \mathrm{d}_{ij}^{\text{ROI}}(a)\right) \\
&= \mathrm{E}\left(\sum_{a \in \mathcal{A}} \sum_{k \neq a} |\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|\right) \\
&= \sum_{a \in \mathcal{A}} \sum_{k \neq a} \mathrm{E}\left(|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|\right) \\
&= \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{2}{\sqrt{\pi}} \\
&= \frac{2p(p-1)}{\sqrt{\pi}}.
\end{aligned} \tag{135}
$$

Due to the dependencies that exist between terms in the double sum when computing 570 the rs-fMRI distance, linearity no longer applies to the variance operator. We proceed 571

by writing the form of the variance as follows

$$
\begin{aligned}
\mathrm{Var}(D_{ij}^{\mathrm{fMRI}}) &= \mathrm{Var}\left(\sum_{a\in\mathcal{A}}\sum_{k\neq a}|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|\right) \\
&= \sum_{a=1}^{p-1}\mathrm{Var}\left(\sum_{k=a+1}^{p} 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|\right) \\
&\quad + 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p} 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^{p} 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}|\right) \\
&= \sum_{a=1}^{p-1}\sum_{k=a+1}^{p}\mathrm{Var}\left(2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|\right) \\
&\quad + 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p} 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^{p} 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}|\right) \qquad (136) \\
&= \sum_{a=1}^{p-1}\sum_{k=a+1}^{p-1}\frac{4(\pi-2)}{\pi} \\
&\quad + 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p} 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^{p} 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}|\right) \\
&= \frac{2p(\pi-2)(p-1)}{\pi} \\
&\quad + 2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p} 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^{p} 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}|\right).
\end{aligned}
$$

In order to have a formula in terms of the number of ROIs p only, we must estimate the double sum on the right-hand side of Eq. 136. Through simulation, it can be seen that the difference between the sample variance $S_{D_{ij}}^2$ and $\frac{2p(\pi-2)(p-1)}{\pi}$ has a quadratic relationship with p. More explicitly, we have the following relationship

$$
S_{D_{ij}^{\mathrm{fMRI}}}^2 - \frac{2p(\pi-2)(p-1)}{\pi} = \beta_1 p^2 + \beta_0 p. \qquad (137)
$$

The coefficient estimates found through least squares fitting are $\beta_0 = -\beta_1 \approx 0.08$. These estimates allow one to infer a functional form for the double sum in the right-hand side of Eq. 136 that is actually proportional to $\frac{2p(\pi-2)(p-1)}{\pi}$. That is, we have the following formula for approximating the double sum

$$
2\sum_{a=1}^{p-1}\sum_{r=a+1}^{p-1}\mathrm{Cov}\left(\sum_{k=a+1}^{p} 2|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|, \sum_{s=r+1}^{p} 2|\hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)}|\right) = \frac{p(\pi-2)(p-1)}{4\pi}. \quad (138)
$$

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

$$
\mathrm{Var}(D_{ij}^{\mathrm{fMRI}}) = \frac{9p(\pi-2)(p-1)}{4\pi}. \qquad (139)
$$

With the mean and variance estimates given by Eqs. 135 and 139, we have the following asymptotic distribution for rs-fMRI distances

$$
D_{ij}^{\mathrm{fMRI}} \dot\sim \mathcal{N}\left(\frac{2p(p-1)}{\sqrt{\pi}}, \frac{9p(\pi-2)(p-1)}{4\pi}\right). \qquad (140)
$$

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

$$D_{ij}^{\text{fMRI}*} = \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{\left| A_{ak}^{(i)} - A_{ak}^{(j)} \right|}{\max(a) - \min(a)}. \tag{141}$$

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

$$\text{E}\left(X_a^{\max} - X_a^{\min}\right) = 2\mu_{\max}^{(1)}(m,p) = 2\left[\frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m(p-1)}\right)} - \Phi^{-1}\left(\frac{1}{m(p-1)}\right) \right]. \tag{142}$$

The variance can be esimated with the following

$$\text{Var}\left(X_a^{\max} - X_a^{\min}\right) = \frac{\pi^2}{6\log[m(p-1)]}. \tag{143}$$

Let $\mu_{D_{ij}^{\text{fMRI}}}$ and $\sigma_{D_{ij}^{\text{fMRI}}}^2$ denote the mean and variance of the rs-fMRI distance distribution given by Eqs. 135 and 139. Using the formulas for the mean and variance of the max-min normalized distance distribution given in Eq. 86, we have the following asymptotic distribution for the max-min normalized rs-fMRI distances

$$D_{ij}^{\text{fMRI}*} \dot{\sim} \mathcal{N}\left(\frac{\mu_{D_{ij}^{\text{fMRI}}}}{2\mu_{\max}^{(1)}(m,p)}, \frac{6\sigma_{D_{ij}^{\text{fMRI}}}^2 \log[m(p-1)]}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m,p)\right]^2 \log[m(p-1)]} \right). \tag{144}$$

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

Substituting the non-normalized mean given by Eq. 135 into Eq. 144 for the mean of the max-min normalized rs-fMRI metric, we have the following

$$\begin{aligned} \text{E}\left(D_{ij}^{\text{fMRI}*}\right) &= \frac{\mu_{D_{ij}^{\text{fMRI}}}}{2\mu_{\max}^{(1)}(m,p)} \\ &= \frac{p(p-1)}{\sqrt{\pi}\mu_{\max}^{(1)}(m,p)}, \end{aligned} \tag{145}$$

where $\mu_{\max}^{(1)}(m,p)$ is given in Eq. 142.

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

$$\begin{aligned} \text{Var}\left(D_{ij}^{\text{fMRI}*}\right) &= \frac{6\sigma_{D_{ij}^{\text{fMRI}}}^2 \log[m(p-1)]}{\pi^2 + 24\left[\mu_{\max}^{(1)}(m,p)\right]^2 \log[m(p-1)]} \\ &= \frac{27(\pi-2)\log[m(p-1)](p-1)p}{2\pi\left(\pi^2 + 24\left[\mu_{\max}^{(1)}(m,p)\right]^2 \log[m(p-1)]\right)}, \end{aligned} \tag{146}$$

where $\mu_{\max}^{(1)}(m,p)$ is given in Eq. 142.

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

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

5 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.

5.1 Continuous data

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

that exists in the simulated data. This measure is given by

$$\bar{r}_{\mathrm{abs}} = \frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j>i} r_{ij} \tag{147}$$

where r_{ij} is the correlation between features i and j across all instances m. The distance density curve (orange) is representative of distances generated from random standard normal data with no added correlation. The mean and variance of this distribution are given by Eqs. 48 and 47, respectively, by substituting $p = 100$ for the mean. From left-to-right and top-to-bottom, there is an increase in $\bar{r}_{\mathrm{abs}}$. This very quickly introduces positive skewness and increased variability. The predicted and sample means, however, are approximately the same in each case due to linearity of the expectation operator. Because of the dependencies between features, the predicted variance of 1 obviously no longer holds.

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}

$$X^{\mathrm{corr}} = XU^{\mathrm{T}}. \tag{148}$$

The new data matrix given by Eq. 148 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.

5.2 GWAS data

In analogy to the previous section, we explore the effects of pairwise feature correlation in the context of GWAS data. Without loss of generality, we let $m = p = 100$ and consider the TiTv metric, which is given by combining Eqs. 106 and 1 with $q = 1$. To create correlated GWAS data, we first generated standard normal data with random correlation structure. We then applied the standard normal cumulative distribution function (CDF) to this correlated data, which was subsequently followed by the application of the inverse binomial CDF with random success probabilities for each feature (or SNP). The resulting GWAS data set is binomial with $n = 2$ trials and has roughly the same correlation matrix as the correlated standard normal data.

5.3 Correlation-based data

For our correlation data-based metric given by Eqs. 134 and 1 with $q = 1$, we consider additional effects of correlation between features. Without loss of generality, we let $m = 100$ and $p = 30$. As in the previous subsections, an illustration of the effects of correlated features in this context is shown in Fig. **??**. Based on the correlated distance densities (blue), it appears that correlation between features introduces positive skewness at lower values of $\bar{r}_{\mathrm{abs}}$. We introduced correlation to the transformed data matrix given by Fig. 6 with the cholesky method used previously.
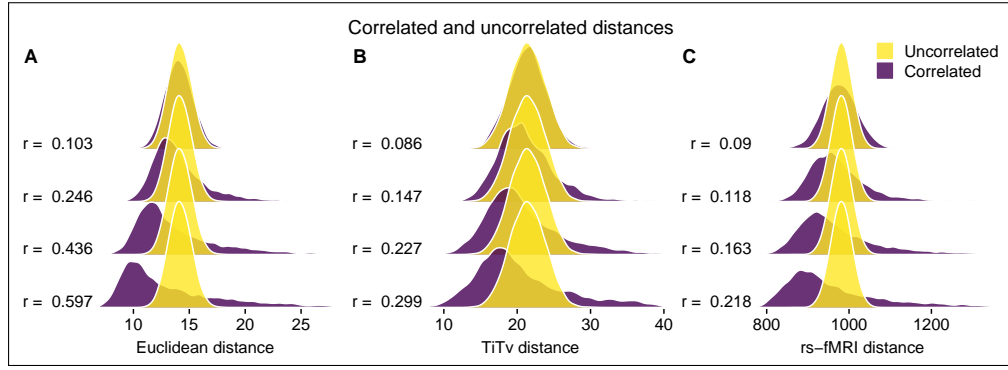
Fig 7. Ridgeline plots of uncorrelated vs correlated densities in bioinformatics data. (**A**)

6 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. 42) and $\sqrt{p}$ (see Eq. 46), 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. 134) 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. 106 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. 104) and AM (see Eq. 105) 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. 118) 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. In Figs. **??** and **??**, we showed how both minor allele frequency and Ti/Tv ratio uniquely affects the TiTv distance. 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. Strong correlations lead to positive skewness in the distance distribution, as shown in Figs. **??**, **??**, and **??**. 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. 147). 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.   ₇₄₆

2. Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and ₇₄₇
  Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal* ₇₄₈
  of Biomedical Informatics, 2018.   ₇₄₉

3. Marko Robnik Šikonja and Igor Kononenko. Theoretical and Empirical Analysis ₇₅₀
  of ReliefF and RReliefF. *Machine Learning*, 53:23 – 69, February 2003. ₇₅₁

4. Archana Venkataraman, Marek Kubicki, Carl-Fredrik Westin, and Polina Golland. ₇₅₂
  Robust Feature Selection in Resting-State fMRI Connectivity Based on Population ₇₅₃
  Studies. *Conf Comput Vis Pattern Recognit Workshops*, pages 63–70, 2010. ₇₅₄

5. Etay Hay, Petra Ritter, Nancy J. Lobaugh, and Anthony R. McIntosh. Mul- ₇₅₅
  tiregional integration in the brain during resting-state fMRI activity. *PLOS* ₇₅₆
  Computational Biology, March 2017.   ₇₅₇

6. Benedikt Sundermann, Mona Olde lütke Beverborg, and Bettina Pfleiderer. Toward ₇₅₈
  literature-based feature selection for diagnostic classification: a meta-analysis of ₇₅₉
  resting-state fMRI in depression. *Frontiers in Human Neuroscience*, September ₇₆₀
  2014.   ₇₆₁

7. Svyatoslav Vergun, Alok S. Deshpande, Timothy B. Meier, Jie Song, Dana ₇₆₂
  L. Tudorascu, Veena A. Nair, Vikas Singh, Bharat B. Biswal, M. Elizabeth ₇₆₃
  Meverand, Rasmus M. Birn, and Vivek Prabhakaran. Characterizing functional ₇₆₄
  connectivity differences in aging adults using machine learning on resting state ₇₆₅
  fMRI data. *Frontiers in Computational Neuroscience*, April 2013. ₇₆₆

8. Trang T Le, Ryan J Urbanowicz, Jason H Moore, and Brett A McKinney. Statis- ₇₆₇
  tical inference relief (stir) feature selection. *Bioinformatics*, page bty788, 2018. ₇₆₈

9. M. Arabnejad, B. A. Dawkins, W. S. Bush, B. C. White, A. R. Harkness, and ₇₆₉
  B. A. McKinney. Transition-transversion encoding and genetic relationship metfic ₇₇₀
  in ReliefF feature selection improves pathway enrichment in GWAS. *BioData* ₇₇₁
  Mining, 11(23), 2018.   ₇₇₂

10. Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* ₇₇₃
  Springer, New York, NY, 2004.   ₇₇₄

11. Irwin Miller and Marylees Miller. *John E. Freund's Mathematical Statistics with* ₇₇₅
  Applications. Pearson Prentice Hall, 7 edition, 2004.   ₇₇₆

12. Alvis Brazma and Jaak Vilo. Gene expression data analysis. *FEBS Letters*, ₇₇₇
  480:17–24, June 2000.   ₇₇₈

13. Dongfang Wang and Jin Gu. VASC: Dimension Reduction and Visualization of ₇₇₉
  Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genomics Proteomics* ₇₈₀
  Bioinformatics, 16:320–331, December 2018.   ₇₈₁

14. Ryan J. Urbanowicz, Melissa Meeker, William LaCava, Randal S. Olson, and ₇₈₂
  Jason H. Moore. Relief-Based Feature Selection: Introduction and Review. ₇₈₃
  arXiv:1711.08421 [cs.DS], 2018.   ₇₈₄

15. E. J. Gumbel. The Distribution of the Range. *The Annals of Mathematical* ₇₈₅
  Statistics, 18(3):384–412, September 1947.   ₇₈₆

16. Sourav Chatterjee. *Superconcentration and Related Topics.* 1439-7382. Springer ₇₈₇
  International Publishing, 1 edition, 2014.   ₇₈₈

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

18. 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.

19. 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.