

# Theoretical properties of distance distributions and novel metrics for nearest-neighbor feature selection

Bryan A. Dawkins<sup>1</sup>, Trang T. Le<sup>2</sup> Brett A. McKinney<sup>3,4\*</sup>,

**1** Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, USA

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

**3** Department of Mathematics, University of Tulsa, Tulsa, OK 74104, USA

**4** Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104, USA

\* brett.mckinney@gmail.com

## Abstract

The performance of nearest-neighbor feature selection and prediction methods depends on the metric for computing neighborhoods and the distribution properties of the underlying data. The effects of the distribution and metric, as well as the presence of correlation between predictors and instances (or samples), are reflected in the expected moments of the distribution of pairwise distances. Much work to introduce new and improved nearest-neighbor feature selection algorithms has been done, which includes new neighborhood methods and distance metrics, but not much attention has been given to the distributional properties of distances as a function of particular data types. We derive general analytical expressions for the mean and variance of pairwise distances for  $L_q$  metrics for normal and uniform random data with  $p$  attributes and  $m$  instances. We use extreme value theory to derive results for metrics that are normalized by the range of each attribute (max - min). In addition to these expressions for continuous data, we derive similar analytical formulas for a new metric for genetic variants (categorical data) in genome-wide association studies (GWAS). The genetic distance distributions account for minor allele frequency and transition/transversion ratio. We introduce a new metric for resting-state functional MRI data (rs-fMRI) and derive its distance properties. This metric is applicable to correlation-based predictors derived from time series data. Derivations assume independent data, which is necessary for defining closed form expressions of moment estimates, but empirically we also consider the effect of correlation between attributes. These analytical results and new metrics can be used to inform the optimization of nearest neighbor methods for a broad range of studies including gene expression, GWAS, and fMRI data. The summary of distribution moments and detailed derivations provide a resource for understanding the distance properties for various metrics and data types.

## Introduction

Statistical models can deviate from expected behavior depending on whether certain properties of the underlying data are satisfied, such as being normally distributed. The expected behavior of nearest neighbor models is further influenced by the choice of metric, such as Euclidean or Manhattan. For random normal data ( $\mathcal{N}(0, 1)$ ), for

example, the variance of the pairwise distances of a Manhattan metric is proportional to the number of attributes ( $p$ ) whereas the variance is constant for a Euclidean metric. Relief methods [1–3] and nearest-neighbor projected-distance regression (NDPR) [4] use nearest neighbors to compute attribute importance scores and often use adaptive neighborhoods that rely on the mean and variance of the distance distribution. The ability of this class of methods to identify association effects, like main effects or interaction effects, depends on parameters such as neighborhood radii or number of neighbors  $k$  [5, 6]. Thus, knowledge of the expected values for a given metric and data distribution may improve the performance of these feature selection methods by informing the selection of neighborhood parameters.

For continuous data, the metrics most commonly used in nearest neighbor methods are  $L_q$  with  $q = 1$  (Manhattan) or  $q = 2$  (Euclidean). For data from standard normal ( $\mathcal{N}(0, 1)$ ) or standard uniform ( $\mathcal{U}(0, 1)$ ) distributions, the asymptotic behavior of the  $L_q$  metrics is known. However, detailed derivations of these distance distribution asymptotics are not readily available in the literature. We provide detailed derivations of generalized expressions parameterized by metric  $q$ , attributes  $p$ , and samples  $m$ , and we extend the derivations to  $L_q$  metrics normalized by the range of the attributes using Extreme Value Theory (EVT). These range (max-min) normalized metrics are often used in Relief-based algorithms [3].

For genome-wide association study (GWAS) data, which is categorical, various metrics have been developed for feature selection and for computing similarity between individuals based on shared genetic variation. We build on the mathematical formalism for continuous data to derive the asymptotic properties of various categorical data metrics for GWAS. We derive asymptotic formulas for the mean and variance for three recently introduced GWAS metrics [7]. These metrics were developed for Relief-based feature selection to account for binary genotype differences (two levels), allelic differences (three levels), and transition/transversion differences (five levels). The mean and variance expressions we derive for these multi-level categorical data types are parameterized by the minor allele frequency and the transition/transversion ratio.

Resting-state fMRI (rs-fMRI) is a growing application area for machine learning and feature selection [8–11], which involves correlation data derived from time-series brain activity. For a given subject, a correlation or similar matrix is computed between brain regions of interest (ROIs) from their time series. Each time series represents functional activity of the ROI while the subject is not performing any task, and the ROI typically corresponds to a region with known function for emotion or cognition. Thus, the dataset consists of pairwise ROI correlations for each of the  $m$  subjects. Nearest-neighbor based feature selection was applied to rs-fMRI in the private evaporative cooling method [12], where the predictors were pairwise correlations between ROIs. The use of pairwise correlation predictors is a common practice because of convenience and differential connectivity between brain regions may be of biological importance [13]. However, one may be interested in the importance of attributes at the ROI level. Thus, in the current study we introduce a new metric to be used in NPDR [4] with resting state correlation matrices that provides attribute importance for ROIs. This metric is applicable to general time series derived correlation data, and we derive asymptotic estimates for the mean and variance of distance distributions for our new time series, correlation based metric.

In Section 1, we introduce preliminary notation and apply the Central Limit Theorem (CLT) and the Delta Method to derive asymptotics for pairwise distances. In Section 2, we present general derivations for continuously distributed data sets with  $m$  instances and  $p$  attributes. We begin with the cases of standard normal ( $\mathcal{N}(0, 1)$ ) and standard uniform ( $\mathcal{U}(0, 1)$ ) data distributions, but we derive analytical expressions parameterized by  $q$ ,  $p$ , and  $m$ . In Section 2.4 we use Extreme Value Theory (EVT) to

derive attribute range-normalized (max-min) versions of  $L_q$  metrics. In Section 3, we extend the derivations to categorical data with a binomial distribution for GWAS data with multiple metric types. In Section 4, we present a new time series correlation-based distance metric, with a particular emphasis on rs-fMRI data, and we derive the corresponding asymptotic distance distribution results. Lastly, in Section 6, we demonstrate the effect of correlation in the attribute space on distance distributional properties.

## 1 Limit distribution for $L_q$ on null data

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

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

which is typically Manhattan ( $q = 1$ ) in Relief-based methods 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 ( $d^{\text{num}}$ ) attribute  $a$  may be

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

where  $\hat{X}$  represents the standardized data matrix  $X$ . We use a simplified  $d_{ij}(a)$  notation in place of the  $\text{diff}(a, (i, j))$  notation that is customary in Relief-based methods. 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  $d_{ij}^{\text{num}}(a)$  projection is simply the absolute difference between row elements  $i$  and  $j$  of the data matrix  $X^{m \times p}$  for the attribute column  $a$ .

### 1.1 Nearest-neighbor projected-distance regression

As in the case of preceding nearest-neighbor feature selection algorithms, the performance of NPDR relies on appropriate choice of neighborhood selection criteria. The size of neighborhoods must be chosen appropriately for optimal detection of important statistical effects. It has been shown using simulations that neighborhood order should be as large as possible to optimally detect main effects, whereas smaller neighborhoods are necessary for detecting interactions [6]. NPDR allows for any neighborhood algorithm to be used, such as fixed or adaptive  $k$ , and fixed or adaptive radius. Especially in the case of radius methods, one needs some sense of central tendency with respect to pairwise distances between a given target instance and its neighbors. Choosing a radius that is too large will potentially yield too many neighbors if the goal of feature selection is finding interaction effects. On the other hand, small radii will inhibit the detection of main effects. A related problem exists for choosing a fixed  $k$  number of neighbors. We would like to choose  $k$  so that the average distance within neighborhoods is not too large or too small with respect to the empirical average pairwise distance between pairs of instances. In order for the appropriate choice of

neighborhood size to be made, we need to know the central tendency and scale of the distance distribution generated on our data.

In the case of dichotomous outcomes, NPDR estimates regression coefficients of the following model

$$\text{logit}(p_{ij}^{\text{miss}}) = \beta_0 + \beta_a d_{ij}(a) + \epsilon_{ij}, \quad (3)$$

where  $p_{ij}^{\text{miss}}$  is the probability of instances  $i, j \in \mathcal{I}$  being in different classes,  $\beta_a$  indicates the relative importance of attribute  $a \in \mathcal{A}$  for predicting the binary outcome, and  $d_{ij}(a)$  is the attribute diff (Eq. 2). The outcome of NPDR, modeled by  $p_{ij}^{\text{miss}}$ , is the diff computed as a function of instance  $i, j \in \mathcal{I}$  class labels, which is given by the following

$$d_{ij}^{\text{miss}}(\vec{y}) = \begin{cases} 0, & y_i = y_j \\ 1, & \text{else,} \end{cases} \quad (4)$$

where  $\vec{y}$  is the binary response (or outcome). The purpose of NPDR is ultimately testing the one-sided hypotheses given by

$$\begin{aligned} H_0 : \beta_a &\leq 0 \\ H_1 : \beta_a &> 0, \end{aligned} \quad (5)$$

where rejecting the null hypothesis ( $H_0$ ) implies that attribute  $a \in \mathcal{A}$  is important for classification.

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 (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 (Eq. 3), however, are the one-dimensional projected distances between two instances  $i, j \in \mathcal{I}$  (Eq. 2). Hence, all asymptotic estimates we derive for diff metrics (Eq. 2) are particularly relevant to NPDR. Since the standard distance metric (Eq. 1) is a function of the one-dimensional projection (Eq. 2), asymptotic estimates derived for this projection (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 standard distance metric (Eq. 1).

## 1.2 Asymptotic normality of pairwise distances

Suppose that  $X_{ia}, X_{ja} \stackrel{iid}{\sim} \mathcal{F}_X(\mu_X, \sigma_X^2)$  for two fixed and distinct instances  $i, j \in \mathcal{I}$  and a fixed attribute  $a \in \mathcal{A}$ .  $\mathcal{F}_X$  represents any data distribution with mean  $\mu_X$  and variance  $\sigma_X^2$ .

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

$$Z_a^q = |d_{ij}(a)|^q = |X_{ia} - X_{ja}|^q, \quad a \in \mathcal{A}. \quad (6)$$

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

$$\left(D_{ij}^{(q)}\right)^q = \sum_{a \in \mathcal{A}} |d_{ij}(a)|^q = \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q = \sum_{a \in \mathcal{A}} Z_a^q \sim \mathcal{N}(\mu_{z_a^q} p, \sigma_{z_a^q}^2 p). \quad (7)$$

Consider the smooth function  $g(z) = z^{1/q}$  that is continuously differentiable for  $z > 0$ . Assuming that  $\mu_{z_a^q} > 0$ , the Delta Method [14] can be applied to show that

$$\begin{aligned}
g\left(\left(D_{ij}^{(q)}\right)^q\right) &= g\left(\sum_{a \in \mathcal{A}} Z_a^q\right) \\
&= \left(\sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q\right)^{1/q} \\
&= D_{ij}^{(q)} \sim \mathcal{N}\left(g\left(\mu_{z_a^q} p\right), \left[g'\left(\mu_{z_a^q} p\right)\right]^2 \sigma_{z_a^q}^2 p\right) \\
\Rightarrow D_{ij}^{(q)} &\sim \mathcal{N}\left(\left(\mu_{z_a^q} p\right)^{1/q}, \frac{\sigma_{z_a^q}^2 p}{q^2 \left(\mu_{z_a^q} p\right)^{2\left(1-\frac{1}{q}\right)}}\right).
\end{aligned} \tag{8}$$

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  $\mathcal{N}\left(\sqrt{\mu_{z_a^2} p}, \frac{\sigma_{z_a^2}^2}{4\mu_{z_a^2}}\right)$ . When  $p$  is small, however, we observe empirically that a closer estimate of the sample mean is

$$\begin{aligned}
\mathbb{E}\left(D_{ij}^{(2)}\right) &= \sqrt{\mathbb{E}\left[\left(D_{ij}^{(2)}\right)^2\right] - \text{Var}\left(D_{ij}^{(2)}\right)} \\
&= \sqrt{\mu_{z_a^2} p - \frac{\sigma_{z_a^2}^2}{4\mu_{z_a^2}}}.
\end{aligned} \tag{9}$$

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

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

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 we have previously given (Eq. 8). 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.

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

For distance based learning methods, all pairwise distances are used to determine relative importances for attributes. The collection of all distances above the diagonal in an  $m \times m$  distance matrix does not satisfy the independence assumption used in the previous derivations. This is because of the redundancy that is inherent to the distance matrix calculation. However, this collection is still asymptotically normal with mean and variance approximately equal to those we have previously given (Eq. 8). 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 (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 ( $L_2$ ) metrics. We then consider the max-min normalized version of the  $L_q$  distance, where the magnitude difference (Eq. 2) is divided by the range of each attribute  $a$ . Using Extreme Value Theory (EVT), we derive formulas for the moments of attribute 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  $X_{ia}, X_{ja} \stackrel{iid}{\sim} \mathcal{F}_X(\mu_x, \sigma_x^2)$  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 [15].

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

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

Elsewhere,  $g(y) = 0$ .

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

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

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

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

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

- (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  $Z_a^q$  is given by

$$\begin{aligned} g^{(2)}(x_{ia}, z_a) &= f_X(x_{ia}, x_{ja}) \left| \frac{\partial x_{ja}}{\partial z_a} \right| \\ &= f_X(x_{ia}) f_X(x_{ja}) \left| \frac{1}{q} (z_a^q)^{\frac{1}{q}-1} \right| \\ &= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} f_X(x_{ia}) f_X \left( x_{ia} + (z_a^q)^{1/q} \right), \quad z_a > 0. \end{aligned} \quad (12)$$

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

$$\begin{aligned} f_{Z_a^q}^{(2)}(z_a^q) &= \int_{-\infty}^{\infty} g^{(2)}(x_{ia}, z_a^q) dx_{ia} \\ &= \frac{1}{q (z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) f_X \left( x_{ia} + (z_a^q)^{1/q} \right) dx_{ia}, \quad z_a > 0. \end{aligned} \quad (13)$$

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

$$E^{(1)} = \{ |X_{ia} - X_{ja}|^q \leq z_a^q : X_{ja} = X_{ia} - (Z_a^q)^{1/q} \} \quad (14)$$

and

$$E^{(2)} = \{ |X_{ia} - X_{ja}|^q \leq z_a^q : X_{ja} = X_{ia} + (Z_a^q)^{1/q} \}. \quad (15)$$

Then it follows from fundamental rules of probability that

213

$$\begin{aligned}
F_{Z_a^q}(z_a^q) &= \mathbb{P}[Z_a^q \leq z_a^q] \\
&= \mathbb{P}[|X_{ia} - X_{ja}|^q \leq z_a^q] \\
&= \mathbb{P}[E^{(1)} \cup E^{(2)}] \\
&= \mathbb{P}[E^{(1)}] + \mathbb{P}[E^{(2)}] - \mathbb{P}[E^{(1)} \cap E^{(2)}] \\
&= \mathbb{P}[E^{(1)}] + \mathbb{P}[E^{(2)}] \\
&= \int_{-\infty}^{z_a^q} f_{Z_a^q}^{(1)}(t) dt + \int_{-\infty}^{z_a^q} f_{Z_a^q}^{(2)}(t) dt \\
&= \int_{-\infty}^{z_a^q} \left( f_{Z_a^q}^{(1)}(t) + f_{Z_a^q}^{(2)}(t) \right) dt \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{z_a^q} \left( \int_{-\infty}^{\infty} f_X(x_{ia}) [f_X(x_{ia} - t) + f_X(x_{ia} + t)] dx_{ia} \right) dt, \quad z_a > 0.
\end{aligned} \tag{16}$$

It follows directly from the previous result (Eq. 16) that the density function of the random variable  $Z_a^q$  is given by

214

215

$$\begin{aligned}
f_{Z_a^q}(z_a^q) &= \frac{\partial}{\partial z_a^q} F_{Z_a^q}(z_a^q) \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[ f_X(x_{ia} - (z_a^q)^{1/q}) + f_X(x_{ia} + (z_a^q)^{1/q}) \right] dx_{ia},
\end{aligned} \tag{17}$$

where  $z_a > 0$ .

216

Using the previous result (Eq. 17), we can compute the mean and variance of the random variable  $Z_a^q$  as

217

218

$$\mu_{z_a^q} = \int_{-\infty}^{\infty} z_a^q f_{Z_a^q}(z_a^q) dz_a^q \tag{18}$$

and

219

$$\sigma_{z_a^q}^2 = \int_{-\infty}^{\infty} (z_a^q)^2 f_{Z_a^q}(z_a^q) dz_a^q - \mu_{z_a^q}^2. \tag{19}$$

It follows immediately from the mean (Eq. 18) and variance (Eq. 19) and the Classical Central Limit Theorem (CCLT) that

220

221

$$\left( D_{ij}^{(q)} \right)^q = \sum_{a \in \mathcal{A}} Z_a^q = \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \sim \mathcal{N}(\mu_{z_a^q} p, \sigma_{z_a^q}^2 p). \tag{20}$$

Applying the convergence result we derived previously (Eq. 8), the distribution of  $D_{ij}^{(q)}$  is given by

222

223

$$D_{ij}^{(q)} \sim \mathcal{N} \left( (\mu_{z_a^q} p)^{1/q}, \frac{\sigma_{z_a^q}^2 p}{q^2 (\mu_{z_a^q} p)^{2(1-\frac{1}{q})}} \right), \quad \mu_{z_a^q} > 0, \tag{21}$$

where we have an improved estimate of the mean for  $q = 2$  (Eq. 9).

224

### 2.1.1 Standard normal data

225

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

226

227

$$f_X(x_{ia}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_{ia}^2}, \tag{22}$$



$$f_X \left( x_{ia} - (z_a^q)^{1/q} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (x_{ia} - (z_a^q)^{1/q})^2}, \quad z_a > 0, \text{ and} \quad (23)$$

$$f_X \left( x_{ia} + (z_a^q)^{1/q} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (x_{ia} + (z_a^q)^{1/q})^2}, \quad z_a > 0. \quad (24)$$

Substituting these marginal densities (Eqs. 22-24) into the general density function for  $Z_a^q$  (Eq. 17) and completing the square on  $x_{ia}$  in the exponents, we have

$$\begin{aligned} f_{Z_a^q}(z_a^q) &= \frac{1}{2q\pi (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \left( e^{-\frac{1}{2} [\sqrt{2}x_{ia} - \frac{\sqrt{2}}{2}(z_a^q)^{1/q}]^2} \right. \\ &\quad \left. + e^{-\frac{1}{2} [\sqrt{2}x_{ia} + \frac{\sqrt{2}}{2}(z_a^q)^{1/q}]^2} \right) dx_{ia} \\ &= \frac{1}{2q\sqrt{\pi} (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left( e^{-\frac{1}{2}u^2} + e^{-\frac{1}{2}u^2} \right) du \\ &= \frac{1}{2q\sqrt{\pi} (z_a^q)^{1-\frac{1}{q}}} e^{-\frac{1}{4}(z_a^q)^{2/q}} (1 + 1) \\ &= \frac{1}{q\sqrt{\pi}} (z_a^q)^{\frac{1}{q}-1} e^{-\frac{1}{4}(z_a^q)^{2/q}} \\ &= \frac{\frac{2}{q}}{(2^q)^{1/q} \Gamma\left(\frac{1}{\frac{2}{q}}\right)} (z_a^q)^{\frac{1}{q}-1} e^{-\left(\frac{z_a^q}{2^q}\right)^{2/q}}. \end{aligned} \quad (25)$$

The density function given previously (Eq. 25) is a Generalized Gamma density with parameters  $b = \frac{2}{q}$ ,  $c = 2^q$ , and  $d = \frac{1}{q}$ . This distribution has mean and variance given by

$$\begin{aligned} \mu_{z_a^q} &= \frac{c\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \\ &= \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} \end{aligned} \quad (26)$$

and

$$\begin{aligned} \sigma_{z_a^q}^2 &= c^2 \left[ \frac{\Gamma\left(\frac{d+2}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} - \left( \frac{\Gamma\left(\frac{d+1}{b}\right)}{\Gamma\left(\frac{d}{b}\right)} \right)^2 \right] \\ &= 4^q \left[ \frac{\Gamma\left(q + \frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2}q + \frac{1}{2}\right)}{\pi} \right]. \end{aligned} \quad (27)$$

By linearity of the expected value and variance operators under the iid assumption, the mean (Eq. 26) and variance (Eq. 27) of the random variable  $Z_a^q$  allow the  $p$ -dimensional mean and variance of the  $\left(D_{ij}^{(q)}\right)^q$  distribution to be computed directly as

$$\mu_{\left(D_{ij}^{(q)}\right)^q} = \mathbb{E} \left[ \left( D_{ij}^{(q)} \right)^q \right] = \mathbb{E} \left( \sum_{a \in \mathcal{A}} Z_a^q \right) = \sum_{a \in \mathcal{A}} \mathbb{E} (Z_a^q) = \sum_{a \in \mathcal{A}} \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} = \frac{2^q\Gamma\left(\frac{q+1}{2}\right)}{\sqrt{\pi}} p \quad (28)$$

and

237

$$\begin{aligned}
\sigma^2_{(D_{ij}^{(q)})^q} &= \text{Var} \left[ \left( D_{ij}^{(q)} \right)^q \right] = \text{Var} \left( \sum_{a \in \mathcal{A}} Z_a^q \right) \\
&= \sum_{a \in \mathcal{A}} \text{Var} (Z_a^q) \\
&= \sum_{a \in \mathcal{A}} 4^q \left[ \frac{\Gamma(q + \frac{1}{2})}{\sqrt{\pi}} - \frac{\Gamma^2(\frac{1}{2}q + \frac{1}{2})}{\pi} \right] \\
&= 4^q \left[ \frac{\Gamma(q + \frac{1}{2})}{\sqrt{\pi}} - \frac{\Gamma^2(\frac{1}{2}q + \frac{1}{2})}{\pi} \right] p.
\end{aligned} \tag{29}$$

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

238

$$\mathcal{N} \left( \left( 2^q \frac{\Gamma(\frac{q+1}{2})}{\sqrt{\pi}} p \right)^{1/q}, \frac{4^q p}{q^2 \left( \frac{2^q \Gamma(\frac{1}{2}q + \frac{1}{2})}{\sqrt{\pi}} p \right)^{2(1-\frac{1}{q})}} \left[ \frac{\Gamma(q + \frac{1}{2})}{\sqrt{\pi}} - \frac{\Gamma^2(\frac{1}{2}q + \frac{1}{2})}{\pi} \right] \right). \tag{30}$$

We provide a summary figure of the moment estimates (Eq. 30) for the  $L_q$  metric on standard normal data (Fig. 2). The summary is organized by data type, type of statistic (mean or variance), and corresponding asymptotic formula. In the next section, we derive the  $L_q$  distance distribution on standard uniform data in a similar fashion.

239

240

241

242

### 2.1.2 Standard uniform data

243

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

244

245

$$f_X(x_{ia}) = 1, \quad 0 \leq x_{ia} \leq 1 \tag{31}$$

246

$$f_X(x_{ia} - (z_a^q)^{1/q}) = 1, \quad 0 \leq x_{ia} - (z_a^q)^{1/q} \leq 1, \text{ and} \tag{32}$$

247

$$f_X(x_{ia} + (z_a^q)^{1/q}) = 1, \quad 0 \leq x_{ia} + (z_a^q)^{1/q} \leq 1. \tag{33}$$

Substituting these marginal densities (Eqs. 31-33) into the more general density function for  $Z_a^q$  (Eq. 17), we have

248

249

$$\begin{aligned}
f_{Z_a^q}(z_a^q) &= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_{-\infty}^{\infty} f_X(x_{ia}) \left[ f_X(x_{ia} - (z_a^q)^{1/q}) + f_X(x_{ia} + (z_a^q)^{1/q}) \right] dx_{ia}, \\
&\quad 0 < z_a \leq 1 \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_0^1 \left[ f_X(x_{ia} - (z_a^q)^{1/q}) + f_X(x_{ia} + (z_a^q)^{1/q}) \right] dx_{ia}, \quad 0 < z_a \leq 1 \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} \int_{(z_a^q)^{1/q}}^1 1 dx_{ia} + \int_0^{1-(z_a^q)^{1/q}} 1 dx_{ia}, \quad 0 < z_a \leq 1 \\
&= \frac{1}{q(z_a^q)^{1-\frac{1}{q}}} [(1 - (z_a^q)^{1/q}) + (1 - (z_a^q)^{1/q})], \quad 0 < z_a \leq 1 \\
&= \frac{1}{q} \cdot 2 (z_a^q)^{\frac{1}{q}-1} [1 - (z_a^q)^{1/q}]^{2-1}, \quad 0 < z_a \leq 1.
\end{aligned} \tag{34}$$

The previous density (Eq. 34) is a Kumaraswamy density with parameters  $b = \frac{1}{q}$  and  $c = 2$  with moment generating function (MGF) given by

$$\begin{aligned} M_n &= \frac{c\Gamma\left(1 + \frac{n}{b}\right)\Gamma(c)}{\Gamma\left(1 + c + \frac{n}{b}\right)} \\ &= \frac{2}{(nq + 2)(nq + 1)}. \end{aligned} \quad (35)$$

Using this MGF (Eq. 35), the mean and variance of  $Z_a^q$  are computed as

$$\mu_{z_a^q} = \frac{2}{(q + 2)(q + 1)} \quad (36)$$

and

$$\sigma_{z_a^q}^2 = \frac{1}{(q + 1)(2q + 1)} - \left( \frac{2}{(q + 2)(q + 1)} \right)^2. \quad (37)$$

By linearity of the expected value and variance operators under the iid assumption, the mean (Eq. 36) and variance (Eq. 37) of the random variable  $Z_a^q$  allow the  $p$ -dimensional mean and variance of the  $\left(D_{ij}^{(q)}\right)^q$  distribution to be computed directly as

$$\begin{aligned} \mu_{\left(D_{ij}^{(q)}\right)^q} &= \mathbb{E} \left[ \left( D_{ij}^{(q)} \right)^q \right] = \mathbb{E} \left( \sum_{a \in \mathcal{A}} Z_a^q \right) \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}(Z_a^q) \\ &= \sum_{a \in \mathcal{A}} \frac{2}{(q + 2)(q + 1)} \\ &= \frac{2p}{(q + 2)(q + 1)} \end{aligned} \quad (38)$$

and

$$\begin{aligned} \sigma_{\left(D_{ij}^{(q)}\right)^q}^2 &= \text{Var} \left[ \left( D_{ij}^{(q)} \right)^q \right] = \text{Var} \left( \sum_{a \in \mathcal{A}} Z_a^q \right) \\ &= \sum_{a \in \mathcal{A}} \text{Var}(Z_a^q) \\ &= \sum_{a \in \mathcal{A}} \left[ \frac{1}{(q + 1)(2q + 1)} - \left( \frac{2}{(q + 2)(q + 1)} \right)^2 \right] \\ &= \left[ \frac{1}{(q + 1)(2q + 1)} - \left( \frac{2}{(q + 2)(q + 1)} \right)^2 \right] p. \end{aligned} \quad (39)$$

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

$$\begin{aligned} \mathcal{N} \left( \left( \frac{2p}{(q + 2)(q + 1)} \right)^{1/q}, \right. \\ \left. \frac{p}{q^2 \left( \frac{2p}{(q + 2)(q + 1)} \right)^{2(1 - \frac{1}{q})}} \left[ \frac{1}{(q + 1)(2q + 1)} - \left( \frac{2}{(q + 2)(q + 1)} \right)^2 \right] \right). \end{aligned} \quad (40)$$

We provide a summary figure of the moment estimates (Eq. 40) for the  $L_q$  metric on standard uniform data (Fig. 2). The summary is organized by data type, type of statistic (mean or variance), and corresponding asymptotic formula. In the next section, we use our general  $L_q$  distance distribution derivations to provide Manhattan ( $q = 1$ ) and Euclidean ( $q = 2$ ) asymptotic moments on both standard normal and standard uniform data. These are the most commonly applied metrics in the context of nearest-neighbor feature selection, so they are of particular interest.

## 2.2 Manhattan ( $L_1$ )

With our general formulas for the asymptotic mean and variance (Eqs. 30 and 40) for any value of  $q \in \mathbb{N}$ , we can simply substitute a particular value of  $q$  in order to determine the asymptotic distribution of the corresponding distance  $L_q$  metric. We demonstrate this with the example of the Manhattan metric ( $L_1$ ) for standard normal and standard uniform data (Eq. 1,  $q = 1$ ).

### 2.2.1 Standard normal data

Substituting  $q = 1$  into the asymptotic formula for the mean  $L_q$  distance (Eq. 30), we have the following for expected  $L_1$  distance between two independently sample instances  $i, j \in \mathcal{I}$  in standard normal data

$$\begin{aligned} \mathbb{E} \left( D_{ij}^{(1)} \right) &= \left( 2 \frac{\Gamma \left( \frac{1+1}{2} \right)}{\sqrt{\pi}} p \right)^{1/1} \\ &= \frac{2p}{\sqrt{\pi}}. \end{aligned} \quad (41)$$

We see in the formula for the expected Manhattan distance (Eq. 41) that  $D_{ij}^{(1)} \sim p$  in the limit, which implies that this distance is unbounded as feature dimension  $p$  increases.

Substituting  $q = 1$  into the formula for the asymptotic variance of  $D_{ij}^{(1)}$  (Eq. 30) leads to the following

$$\begin{aligned} \text{Var} \left( D_{ij}^{(1)} \right) &= \frac{4^1 p}{1^2 \left( \frac{2^1 \Gamma \left( \frac{1}{2}(1) + \frac{1}{2} \right)}{\sqrt{\pi}} p \right)^{2(1 - \frac{1}{1})}} \left[ \frac{\Gamma \left( 1 + \frac{1}{2} \right)}{\sqrt{\pi}} - \frac{\Gamma^2 \left( \frac{1}{2}(1) + \frac{1}{2} \right)}{\pi} \right] \\ &= \frac{2(\pi - 2)p}{\pi}. \end{aligned} \quad (42)$$

Similar to the mean (Eq. 41), the limiting variance of  $D_{ij}^{(1)}$  (Eq. 42) grows on the order of feature dimension  $p$ , which implies that points become more dispersed as the dimension increases. The moment estimates given in this section (Eqs. 41 and 42) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 3).

### 2.2.2 Standard uniform data

Substituting  $q = 1$  into the asymptotic formula of the mean (Eq. 40), we have the following for the expected  $L_1$  distance between two independently sampled instances

$i, j \in \mathcal{I}$  in standard uniform data

289

$$\begin{aligned} \mathbb{E}\left(D_{ij}^{(1)}\right) &= \left(\frac{2p}{(1+2)(1+1)}\right)^{1/1} \\ &= \frac{p}{3}. \end{aligned} \quad (43)$$

Once again, we see that the mean of  $D_{ij}^{(1)}$  (Eq. 43) grows on the order of  $p$  just as in the case of standard normal data.

290  
291

Substituting  $q = 1$  into the formula of the asymptotic variance of  $D_{ij}^{(1)}$  (Eq. 40) leads to the following

292  
293

$$\begin{aligned} \text{Var}\left(D_{ij}^{(1)}\right) &= \frac{p}{1^2 \left(\frac{2p}{(1+2)(1+1)}\right)^{2(1-\frac{1}{1})}} \left[ \frac{1}{(1+1)(2(1)+1)} - \left(\frac{2}{(1+2)(1+1)}\right)^2 \right] \\ &= \frac{p}{18}. \end{aligned} \quad (44)$$

As in the case of the  $L_1$  metric on standard normal data, we have a variance (Eq. 44) that grows on the order of  $p$ . The distances between points in high-dimensional uniform data become more widely dispersed with this metric. The moment estimates given in this section (Eqs. 43 and 44) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 3).

294  
295  
296  
297  
298

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

299  
300

In nearest-neighbor distance-based feature selection like NPDR and Relief-based algorithms, the one-dimensional projection of the pairwise distance onto an attribute (Eq. 2) is particularly fundamental to feature quality for association with an outcome. For instance, this distance projection is the predictor used to determine beta coefficients in NPDR. In particular, understanding distributional properties of the projected distances is necessary for defining pseudo P values for NPDR. In this section, we summarize the exact distribution of the one-dimensional projected distance onto an attribute  $a \in \mathcal{A}$ . These results apply to continuous data, such as gene expression.

301  
302  
303  
304  
305  
306  
307  
308

In previous sections, we derived the exact density function (Eq. 17) and moments (Eqs. 18 and 19) for the distribution of  $Z_a^q = |X_{ia} - X_{ja}|^q$ . We then derived the exact density (Eq. 25) and moments (Eqs. 26 and 27) for standard normal data. Analogously, we formulated the exact density (Eq. 34) and moments (Eqs. 36 and 37) for standard uniform data. From these exact densities and moments, we simply substitute  $q = 1$  to define the distribution of the one-dimensional projected distance onto an attribute  $a \in \mathcal{A}$ .

309  
310  
311  
312  
313  
314  
315

Assuming data is standard normal, we substitute  $q = 1$  into the density function of  $Z_a^q$  (Eq. 25) to arrive at the following density function

316  
317

$$\begin{aligned} f_{Z_a^1}(z_a^1) &= \frac{\frac{2}{1}}{(2^1)^{1/1} \Gamma\left(\frac{1}{2}\right)} (z_a^1)^{1/1-1} e^{-\left(\frac{z_a^1}{2^1}\right)^{2/1}}, \quad z_a > 0 \\ &= \frac{1}{\sqrt{\pi}} z_a e^{-\frac{1}{4}z_a^2}, \quad z_a > 0. \end{aligned} \quad (45)$$

The mean corresponding to this Generalized Gamma density is computed by

318

substituting  $q = 1$  into the formula for the mean of  $Z_a^q$  (Eq. 26). This result is given by 319

$$\begin{aligned}\mu_{Z_a^1} &= \frac{2^1 \Gamma\left(\frac{1+1}{2}\right)}{\sqrt{\pi}} \\ &= \frac{2}{\sqrt{\pi}}.\end{aligned}\tag{46}$$

Substituting  $q = 1$  into Eq. 27 for the variance, we have the following 320

$$\begin{aligned}\sigma_{Z_a^1}^2 &= 4^1 \left[ \frac{\Gamma\left(1 + \frac{1}{2}\right)}{\sqrt{\pi}} - \frac{\Gamma^2\left(\frac{1}{2} \cdot 1 + \frac{1}{2}\right)}{\pi} \right] \\ &= \frac{2(\pi - 2)}{\pi}.\end{aligned}\tag{47}$$

These last few results (Eqs. 45-47) provide us with the distribution for NPDR predictors when the data is from the standard normal distribution. 321

If we have standard uniform data, we substitute  $q = 1$  into the density function of  $Z_a^q$  (Eq. 34) to obtain the following density function 322  
323  
324

$$\begin{aligned}f_{Z_a^1} &= \frac{1}{1} \cdot 2 \left(z_a^1\right)^{1/1-1} \left[1 - \left(z_a^1\right)^{1/1}\right]^{2-1}, \quad 0 < z_a \leq 1 \\ &= 2z_a(1 - z_a), \quad 0 < z_a \leq 1.\end{aligned}\tag{48}$$

The mean corresponding to this Kumaraswamy density is computed by substituting  $q = 1$  into the formula for the mean of  $Z_a^q$  (Eq. 36). After substitution, we have the following result 325  
326  
327

$$\begin{aligned}\mu_{Z_a^1} &= \frac{2}{(1+2)(1+1)} \\ &= \frac{1}{3}.\end{aligned}\tag{49}$$

Substituting  $q = 1$  into the formula for the variance of  $Z_a^q$  (Eq. 37), we have the following 328  
329

$$\begin{aligned}\sigma_{Z_a^1}^2 &= \frac{1}{(1+1)(2 \cdot 1 + 1)} - \left( \frac{2}{(1+2)(1+1)} \right)^2 \\ &= \frac{1}{18}.\end{aligned}\tag{50}$$

In the event that the data distribution is standard uniform, the density function (Eq. 48), the mean (Eq. 49), and the variance (Eq. 50) sufficiently define the distribution for NPDR predictors. The means (Eqs. 46 and 49) and variances (Eqs. 47 and 50) come from the exact distribution of pairwise distances with respect to a single attribute  $a \in \mathcal{A}$ . This is the distribution of the so-called “projection” of the pairwise distance onto a single attribute to which we have been referring, which is a direct implication from our more general derivations. In a similar manner, one can substitute any value of  $q \geq 2$  into the general densities of  $Z_a^q$  for standard normal (Eq. 25) and standard uniform (Eq. 34) to derive the associated density of  $Z_a^q = |X_{ia} - X_{ja}|^q$  for the given data type. 330  
331  
332  
333  
334  
335  
336  
337  
338

### 2.3 Euclidean ( $L_2$ ) 339

Moment estimates for the Euclidean metric are obtained by substituting  $q = 2$  into the asymptotic moment formulas for standard normal data (Eq. 30) and standard uniform data (Eq. 40). As in the case of the Manhattan metric in the previous sections, we initially proceed by deriving Euclidean distance moments in standard normal data. 340  
341  
342  
343

### 2.3.1 Standard normal data

Substituting  $q = 2$  into the asymptotic formula of the mean (Eq. 30), we have the following for expected  $L_2$  distance between two independently sampled instances  $i, j \in \mathcal{I}$  in standard normal data

$$\begin{aligned} \mathbb{E} \left( D_{ij}^{(2)} \right) &= \left( 2 \frac{\Gamma \left( \frac{2+1}{2} \right)}{\sqrt{\pi}} p \right)^{1/2} \\ &= \sqrt{2p}. \end{aligned} \quad (51)$$

In the case of  $L_2$  on standard normal data, we see that the mean of  $D_{ij}^{(2)}$  (Eq. 51) grows on the order of  $\sqrt{p}$ . Hence, the Euclidean distance does not increase as quickly as the Manhattan distance on standard normal data.

Substituting  $q = 2$  into the formula for the asymptotic variance of  $D_{ij}^{(2)}$  (Eq. 30) leads to the following

$$\begin{aligned} \text{Var} \left( D_{ij}^{(2)} \right) &= \frac{4^2 p}{2^2 \left( \frac{2^2 \Gamma \left( \frac{1}{2} (2) + \frac{1}{2} \right)}{\sqrt{\pi}} p \right)^{2(1-\frac{1}{2})}} \left[ \frac{\Gamma \left( 2 + \frac{1}{2} \right)}{\sqrt{\pi}} - \frac{\Gamma^2 \left( \frac{1}{2} (2) + \frac{1}{2} \right)}{\pi} \right] \\ &= 1. \end{aligned} \quad (52)$$

Surprisingly, the asymptotic variance (Eq. 52) is just 1. Regardless of data dimensions  $m$  and  $p$ , the variance of Euclidean distances on standard normal data tends to 1. Therefore, most instances are contained within a ball of radius 1 about the mean in high feature dimension  $p$ . This means that the Euclidean distance distribution on standard normal data is simply a horizontal shift to the right of the standard normal distribution.

For the case in which the number of attributes  $p$  is small, we have an improved estimate of the mean (Eq. 9). The lower dimensional estimate of the mean is given by

$$\begin{aligned} \mathbb{E} \left( D_{ij}^{(2)} \right) &= \left( 2 \frac{\Gamma \left( \frac{2+1}{2} \right)}{\sqrt{\pi}} p - 1 \right)^{1/2} \\ &= \sqrt{2p - 1}. \end{aligned} \quad (53)$$

For high dimensional data sets like gene expression [16, 17], which typically contain thousands of genes (or features), it is clear that the magnitude of  $p$  will be sufficient to use the standard asymptotic estimate (Eq. 51) since  $\sqrt{2p} \approx \sqrt{2p - 1}$  in that case. The moment estimates given in this section (Eqs. 53 and 52) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 3).

### 2.3.2 Standard uniform data

Substituting  $q = 2$  into the asymptotic formula of the mean (Eq. 40), we have the following for expected  $L_2$  distance between two independently sampled instances  $i, j \in \mathcal{I}$  in standard uniform data

$$\begin{aligned} \mathbb{E} \left( D_{ij}^{(2)} \right) &= \left( \frac{2p}{(2+2)(2+1)} \right)^{1/2} \\ &= \sqrt{\frac{p}{6}}. \end{aligned} \quad (54)$$

As in the case of standard normal data, the expected value of  $D_{ij}^{(2)}$  (Eq. 54) grows on the order of  $\sqrt{p}$ .

Substituting  $q = 2$  into the formula for the asymptotic variance of  $D_{ij}^{(2)}$  (Eq. 40) leads to the following

$$\begin{aligned} \text{Var}\left(D_{ij}^{(2)}\right) &= \frac{p}{2^2 \left(\frac{2p}{(2+2)(2+1)}\right)^{2(1-\frac{1}{2})}} \left[ \frac{1}{(2+1)(2(2)+1)} - \left(\frac{2}{(2+2)(2+1)}\right)^2 \right] \\ &= \frac{7}{120}. \end{aligned} \quad (55)$$

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

For the case in which the number of attributes  $p$  is small, we have an improved estimate of the mean (Eq. 9). The lower dimensional estimate of the mean is given by

$$\begin{aligned} \mathbb{E}\left(D_{ij}^{(2)}\right) &= \left(\frac{2p}{(2+2)(2+1)} - \frac{7}{120}\right)^{1/2} \\ &= \sqrt{\frac{p}{6} - \frac{7}{120}}. \end{aligned} \quad (56)$$

We summarize the moment estimates given in this section for standard  $L_q$  metrics (Eqs. 56 and 55) organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 3). In the next section, we extend these results for the standard  $L_q$  metric to derive asymptotics for the attribute range-normalized (max-min)  $L_q$  metric used frequently in Relief-based algorithms [1, 3] for scoring attributes. These derivations use extreme value theory to handle the maximum and minimum attributes for standard normal and standard uniform data.

## 2.4 Distribution of max-min normalized $L_q$ metric

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

$$d_{ij}^{\text{num}}(a) = \text{diff}(a, (i, j)) = \frac{|X_{ia} - X_{ja}|}{\max(a) - \min(a)}, \quad (57)$$

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

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

$$\begin{aligned} D_{ij}^{(q*)} &= \left( \sum_{a \in \mathcal{A}} |d_{ij}(a)|^q \right)^{1/q} \\ &= \left( \sum_{a \in \mathcal{A}} \left( \frac{|X_{ia} - X_{ja}|}{\max(a) - \min(a)} \right)^q \right)^{1/q}. \end{aligned} \quad (58)$$

In order to determine moments of asymptotic max-min normalized distance (Eq. 57) distributions, we will first derive the asymptotic extreme value distributions of the attribute maximum and minimum. Although the exact distribution of the maximum or minimum requires an assumption about the data distribution, the Fisher-Tippett-Gnedenko Theorem is an important result that allows one to generally categorize the extreme value distribution for a collection of independent and identically distributed random variables into one of three distributional families. This theorem does not, however, tell us the exact distribution of the maximum that we require in



order to determine asymptotic results for the max-min normalized distance (Eq. 58).  
 We mention this theorem simply to provide some background on convergence of extreme  
 values. Before stating the theorem, we first need the following definition

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

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

**Theorem 2.2 (Fisher-Tippett-Gnedenko)** Let  $X_{1a}, X_{2a}, \dots, X_{ma} \stackrel{iid}{\sim} \mathcal{F}_X(\mu_x, \sigma_x^2)$   
 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

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

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

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

$$\begin{aligned} P[X_a^{max} \leq x] &= P\left[\max_{k \in \mathcal{I}} \{X_{ka}\} \leq x\right] \\ &= P[X_{1a} \leq x, X_{2a} \leq x, \dots, X_{ma} \leq x] \\ &= \prod_{k=1}^m P[X_{ka} \leq x] \\ &= \prod_{k=1}^m F_X(x) \\ &= [F_X(x)]^m. \end{aligned} \tag{59}$$

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

$$F_{\max}(x) = [F_X(x)]^m, \tag{60}$$

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

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

$$\begin{aligned} f_{\max}(x) &= \frac{d}{dx} F_{\max}(x) \\ &= \frac{d}{dx} [F_X(x)]^m \\ &= m[F_X(x)]^{m-1} f_X(x), \end{aligned} \quad (61)$$

where  $m$  is the size of the sample from which the maximum is derived,  $F_X$  is the distribution function corresponding to the data sample, and  $f_X$  is the density function corresponding to the data sample. Similar to the distribution function for the sample maximum (Eq. 60), the density function (Eq 61) relies only on the distribution and density function of the data from which extremes are derived.

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

$$\begin{aligned} P[X_a^{\min} \leq x] &= 1 - P[X_a^{\min} \geq x] \\ &= 1 - P\left[\min_{k \in \mathcal{I}} \{X_{ka}\} \geq x\right] \\ &= 1 - P[X_{1a} \geq x, X_{2a} \geq x, \dots, X_{ma} \geq x] \\ &= 1 - \prod_{k=1}^m P[X_{ka} \geq x] \\ &= 1 - [P[X_{1a} \geq x]]^m \\ &= 1 - [1 - P[X_{1a} \leq x]]^m \\ &= 1 - [1 - F_X(x)]^m, \end{aligned} \quad (62)$$

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

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

$$F_{\min}(x) = 1 - [1 - F_X(x)]^m. \quad (63)$$

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

Differentiating the distribution function (Eq. 63) gives us the following density function for the distribution of sample minimum

$$\begin{aligned} f_{\min}(x) &= \frac{d}{dx} F_{\min}(x) \\ &= \frac{d}{dx} (1 - [1 - F_X(x)]^m) \\ &= m [1 - F_X(x)]^{m-1} f_X(x), \end{aligned} \quad (64)$$

where  $m$  is the size of the sample from which the minimum is derived,  $F_X$  is the distribution function corresponding to the data sample, and  $f_X$  is the density function corresponding to the data sample. As in the case of the density function for sample maximum (Eq. 61), the density function for sample minimum relies only on the distribution  $F_X$  and density  $f_X$  functions of the data from which extremes are derived and the sample size  $m$ .

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

$$\begin{aligned}\mu_{\max}^{(1)}(m) &= E(X_a^{\max}) = \int_{-\infty}^{\infty} x f_{\max}(x) dx \\ &= \int_{-\infty}^{\infty} x (m[F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x f_X(x) [F_X(x)]^{m-1} dx,\end{aligned}\tag{65}$$

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

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

$$\begin{aligned}\mu_{\max}^{(2)}(m) &= E[(X_a^{\max})^2] = \int_{-\infty}^{\infty} x^2 f_{\max}(x) dx \\ &= \int_{-\infty}^{\infty} x^2 (m[F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x^2 f_X(x) [F_X(x)]^{m-1} dx\end{aligned}\tag{66}$$

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

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

$$\sigma_{\max}^2(m) = \mu_{\max}^{(2)}(m) - \left[\mu_{\max}^{(1)}(m)\right]^2,\tag{67}$$

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

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

$$\begin{aligned}\mu_{\min}^{(1)}(m) &= E(X_a^{\min}) = \int_{-\infty}^{\infty} x f_{\min}(x) dx \\ &= \int_{-\infty}^{\infty} x (m[1 - F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x f_X(x) [1 - F_X(x)]^{m-1} dx,\end{aligned}\tag{68}$$

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

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

$$\begin{aligned}\mu_{\min}^{(2)}(m) &= E[(X_a^{\min})^2] = \int_{-\infty}^{\infty} x^2 f_{\min}(x) dx \\ &= \int_{-\infty}^{\infty} x^2 (m[1 - F_X(x)]^{m-1} f_X(x)) dx \\ &= m \int_{-\infty}^{\infty} x^2 f_X(x) [1 - F_X(x)]^{m-1} dx,\end{aligned}\tag{69}$$

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

Using the first (Eq. 68) and second (Eq. 69) raw moments of the distribution of sample minimum, the variance is given by

$$\sigma_{\min}^2(m) = \mu_{\min}^{(2)}(m) - \left[ \mu_{\min}^{(1)}(m) \right]^2, \quad (70)$$

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

Using the expected attribute maximum (Eq. 65) and minimum (Eq. 68) for sample size  $m$ , the following expected attribute range results from linearity of the expectation operator

$$\begin{aligned} E(X_a^{\max} - X_a^{\min}) &= E(X_a^{\max}) - E(X_a^{\min}) \\ &= \mu_{\max}^{(1)}(m) - \mu_{\min}^{(1)}(m). \end{aligned} \quad (71)$$

where  $\mu_{\max}^{(1)}(m)$  is the expected sample maximum (Eq. 65) and  $\mu_{\min}^{(1)}(m)$  is the expected sample minimum.

For a data distribution whose density is an even function, the expected attribute range (Eq. 71) can be simplified to the following expression

$$E(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m), \quad (72)$$

where  $m$  is the size of the sample from which the maximum is derived. Hence, the expected attribute range is simply twice the expected attribute maximum (Eq. 65). This result naturally applies to standard normal data, which is symmetric about its mean at 0 and without any skewness.

For large samples ( $m \gg 1$ ) from an exponential type distribution that has infinite support and all moments, the covariance between the sample maximum and minimum is approximately zero [18]. In this case, the variance of the attribute range of a sample of size  $m$  is given by the following

$$\begin{aligned} \text{Var}(X_a^{\max} - X_a^{\min}) &\approx \text{Var}(X_a^{\max}) + \text{Var}(X_a^{\min}) \\ &= \sigma_{\max}^2(m) + \sigma_{\min}^2(m). \end{aligned} \quad (73)$$

Under the assumption of zero skewness, infinite support and even density function, sufficiently large sample size  $m$ , and distribution of an exponential type for all moments, the variance of attribute range (Eq. 73) simplifies to the following

$$\begin{aligned} \text{Var}(X_a^{\max} - X_a^{\min}) &= 2\text{Var}(X_a^{\max}) \\ &= 2\sigma_{\max}^2. \end{aligned} \quad (74)$$

Let  $\mu_{D_{ij}^{(q)}}$  and  $\sigma_{D_{ij}^{(q)}}^2$  (Eq. 21) denote the mean and variance of the standard  $L_q$  distance metric (Eq. 1). Then the expected value of the max-min normalized distance

(Eq. 58) distribution is given by the following

$$\begin{aligned}
\mu_{D_{ij}^{(q*)}} &= \mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} \left( \frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{1/q} \right] \\
&\approx \frac{1}{\mathbb{E}(X_a^{\max} - X_a^{\min})} \mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \right)^{1/q} \right] \\
&= \frac{\mu_{D_{ij}^{(q)}}}{\mathbb{E}(X_a^{\max}) - \mathbb{E}(X_a^{\min})} \\
&= \frac{\mu_{D_{ij}^{(q)}}}{\mu_{\max}^{(1)}(m) - \mu_{\min}^{(1)}(m)},
\end{aligned} \tag{75}$$

where  $m$  is the size of the sample from which extremes are derived,  $\mu_{\max}^{(1)}(m)$  is the expected value of the sample maximum (Eq. 65), and  $\mu_{\min}^{(1)}$  is the expected value of the sample minimum.

The variance of the max-min normalized distance (Eq. 58) distribution is given by the following

$$\begin{aligned}
\sigma_{D_{ij}^{(q*)}}^2 &= \text{Var} \left[ \left( \sum_{a \in \mathcal{A}} \left( \frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{1/q} \right] \\
&= \mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} \left( \frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{2/q} \right] - \left( \mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} \left( \frac{|X_{ia} - X_{ja}|}{X_a^{\max} - X_a^{\min}} \right)^q \right)^{1/q} \right] \right)^2 \\
&\approx \frac{\mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \right)^{2/q} \right]}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} - \frac{\left( \mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} |X_{ia} - X_{ja}|^q \right)^{1/q} \right] \right)^2}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} \\
&= \frac{\sigma_{D_{ij}^{(q)}}^2 + \mu_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} - \frac{\mu_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} \\
&= \frac{\sigma_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max} - X_a^{\min})^2]} \\
&= \frac{\sigma_{D_{ij}^{(q)}}^2}{\mathbb{E}[(X_a^{\max})^2] - 2\mathbb{E}(X_a^{\max})\mathbb{E}(X_a^{\min}) + \mathbb{E}(X_a^{\min})^2} \\
&= \frac{\sigma_{D_{ij}^{(q)}}^2}{\mu_{\max}^{(2)}(m) - 2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m) + \mu_{\min}^{(2)}(m)},
\end{aligned} \tag{76}$$

where  $m$  is the size of the sample from which extremes are derived,  $\mu_{\max}^{(1)}(m)$  is the expected value of the sample maximum (Eq. 65), and  $\mu_{\min}^{(1)}$  is the expected value of the sample minimum.

With the mean (Eq. 75) and variance (Eq. 76) of the max-min normalized distance (Eq. 58), we have the following generalized estimate for the asymptotic distribution of the max-min normalized distance distribution

$$D_{ij}^{(q*)} \sim \mathcal{N} \left( \frac{\mu_{D_{ij}^{(q)}}}{\mu_{\max}^{(1)}(m) - \mu_{\min}^{(1)}(m)}, \frac{\sigma_{D_{ij}^{(q)}}^2}{\mu_{\max}^{(2)}(m) - 2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m) + \mu_{\min}^{(2)}(m)} \right), \tag{77}$$

where  $m$  is the size of the sample from which extremes are derived,  $\mu_{\max}^{(1)}(m)$  is the expected value of the sample maximum (Eq. 65), and  $\mu_{\min}^{(1)}$  is the expected value of the sample minimum.

For data with zero skewness and support that is symmetric about 0, the expected sample maximum is the additive inverse of the expected sample minimum. This allows us to express the expected max-min normalized pairwise distance (Eq. 75) exclusively in terms of the expected sample maximum. This result is given by the following

$$\mu_{D_{ij}^{(q*)}} \approx \frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}, \quad (78)$$

where  $m$  is the size of the sample from which the maximum is derived and  $\mu_{\max}^{(1)}(m)$  is the expected value of the sample maximum (Eq. 65).

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

$$\begin{aligned} \sigma_{D_{ij}^{(q*)}}^2 &\approx \frac{\sigma_{D_{ij}^{(q)}}^2}{2\mu_{\max}^{(2)}(m) + 2\left[\mu_{\max}^{(1)}(m)\right]^2} \\ &= \frac{\sigma_{D_{ij}^{(q)}}^2}{2\left(\sigma_{\max}^2(m) + \left[\mu_{\max}^{(1)}(m)\right]^2\right)}, \end{aligned} \quad (79)$$

where  $m$  is the size of the sample from which extremes are derived,  $\mu_{\max}^{(1)}(m)$  is the expected value of the sample maximum (Eq. 65), and  $\sigma_{\max}^2(m)$  is the variance of the sample maximum (Eq. 67).

Therefore, the asymptotic distribution of the max-min normalized distance distribution (Eq. 77) becomes

$$D_{ij}^{(q*)} \sim \mathcal{N}\left(\frac{\mu_{D_{ij}^{(q)}}}{2\mu_{\max}^{(1)}(m)}, \frac{\sigma_{D_{ij}^{(q)}}^2}{2\left(\sigma_{\max}^2(m) + \left[\mu_{\max}^{(1)}(m)\right]^2\right)}\right), \quad (80)$$

where  $m$  is the size of the sample from which extremes are derived,  $\mu_{\max}^{(1)}(m)$  is the expected value of the sample maximum (Eq. 65), and  $\sigma_{\max}^2(m)$  is the variance of the sample maximum (Eq. 67).

We have now derived asymptotic estimates of the moments of the max-min normalized  $L_q$  distance metric (Eq. 58) for any continuous data distribution. In the next two sections, we examine the max-min normalized  $L_q$  distance on standard normal and standard uniform data. As in previous sections in which we analyzed the standard  $L_q$  metric (Eq. 1), we will use the more general results for the max-min  $L_q$  metric to derive asymptotic estimates for normalized Manhattan ( $q = 1$ ) and Euclidean ( $q = 2$ ).

#### 2.4.1 Standard normal data

The standard normal distribution has zero skewness, even density function, infinite support, and all moments. This implies that the corresponding mean and variance of the distribution of sample range can be expressed exclusively in terms of the sample maximum. Given the nature of the density function of the sample maximum for sample size  $m$ , the integration required to determine the moments (Eqs. 65 and 66) is not possible. These moments can either be approximated numerically or we can use extreme

value theory to determine the form of the asymptotic distribution of the sample maximum. Using the latter method, we will show that the asymptotic distribution of the sample maximum for standard normal data is in the Gumbel family. Let  $c_m = -\Phi^{-1}\left(\frac{1}{m}\right)$  and  $b_m = \frac{1}{c_m}$ , where  $\Phi$  is the standard normal cumulative distribution function. Using Taylor's Theorem, we have the following expansion

$$\begin{aligned}\log\Phi(-c_m - b_mx) &= \log\Phi(-c_m) - b_mx \frac{\phi(-c_m)}{\Phi(-c_m)} + \mathcal{O}(b_m^2 x^2) \\ &= \log\left(\frac{1}{m}\right) - x \frac{\phi(-c_m)}{c_m \Phi(-c_m)} + \mathcal{O}(b_m^2 x^2),\end{aligned}\tag{81}$$

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

In order to simplify the right-hand side of this expansion (Eq. 81), we will use the well known Mills Ratio Bounds [19] given by the following

$$1 \leq \frac{\phi(x)}{x\Phi(-x)} \leq 1 + \frac{1}{x^2} \quad , x > 0,\tag{82}$$

where  $\Phi$  and  $\phi$  once again represent the cumulative distribution function and density function, respectively, of the standard normal distribution.

The inequalities given above (Eq. 82) show that

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

This further implies that

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

since

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

This gives us the following approximation of the right-hand side of the expansion (Eq. 81) given previously

$$\begin{aligned}\log\Phi(-c_m - b_mx) &\approx \log\left(\frac{1}{m}\right) - x + \mathcal{O}(b_m^2 x^2) \\ \Rightarrow \Phi(-c_m - b_mx) &\approx \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)} \\ \Rightarrow \Phi(c_m + b_mx) &\approx 1 - \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)},\end{aligned}\tag{83}$$

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

Using the approximation of expansion given previously (Eq. 83), we now derive the

limit distribution for the sample maximum in standard normal data as

557

$$\begin{aligned}
P\left(\frac{X_a^{\max} - c_m}{b_m} \leq x\right) &= P(X_a^{\max} \leq c_m + b_m x) \\
&= \Phi^m(c_m + b_m x) \\
&\approx \left(1 - \frac{1}{m} e^{-x + \mathcal{O}(b_m^2 x^2)}\right)^m \\
&= \left(1 - \frac{1}{m} e^{-x + \mathcal{O}\left(\frac{1}{c_m^2} x^2\right)}\right)^m \\
&\approx \left(1 - \frac{1}{m} e^{-x}\right)^m \\
\Rightarrow \lim_{m \rightarrow \infty} P\left(\frac{X_a^{\max} - c_m}{b_m} \leq x\right) &= \lim_{m \rightarrow \infty} \left(1 - \frac{1}{m} e^{-x}\right)^m \\
&= e^{-e^{-x}},
\end{aligned} \tag{84}$$

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

558  
559

$$E(X_a^{\max}) = \mu_{\max}^{(1)} = -\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)}, \tag{85}$$

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

560  
561  
562

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

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

563

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

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

564  
565

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

566

$$\tilde{\mu}_{\max} = \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right), \tag{86}$$

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

567

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

568  
569

$$\text{Var}(X_a^{\max}) = \frac{\pi^2}{6} \left( \frac{1}{-\Phi^{-1}\left(\frac{1}{m}\right)} \right)^2, \tag{87}$$

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

570

For typical sample sizes  $m$  in high-dimensional spaces, the variance estimate (Eq. 87) exceeds the variance of the sample maximum significantly. Using the fact that

571  
572

$$-\Phi^{-1}\left(\frac{1}{m}\right) \sim \sqrt{2\log(m)} \text{ [20]}$$



and

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

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

$$\begin{aligned} \sigma_{\max}^2(m) = \text{Var}(X_a^{\max}) &\approx \frac{\pi^2}{6} \left( \frac{1}{\sqrt{2\log(m)}} \right)^2 \\ &= \frac{\pi^2}{12\log(m)}. \end{aligned} \quad (88)$$

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

$$\mathbb{E}(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m) = 2 \left[ -\Phi^{-1}\left(\frac{1}{m}\right) - \frac{\gamma}{\Phi^{-1}\left(\frac{1}{m}\right)} \right], \quad (89)$$

where  $\gamma$  is the Euler-Mascheroni constant.

It is well known that the sample extremes from the standard normal distribution are approximately uncorrelated for large sample size  $m$  [18]. This implies that we can approximate the variance of the range of  $m$  iid standard normal random variables with the following result

$$\begin{aligned} \text{Var}(X_a^{\max} - X_a^{\min}) &\approx \text{Var}(X_a^{\max}) + \text{Var}(X_a^{\min}) \\ &= \sigma_{\max}^2(m) + \sigma_{\min}^2(m) \\ &= 2\sigma_{\max}^2(m) \\ &\approx 2 \left( \frac{\pi^2}{12\log(m)} \right) \\ &= \frac{\pi^2}{6\log(m)}. \end{aligned} \quad (90)$$

For the purpose of approximating the mean and variance of the max-min normalized distance distribution, we observe empirically that the formula for the median of the distribution of the attribute maximum (Eq. 86) yields more accurate results. More precisely, the approximation of the expected maximum (Eq. 85) overestimates the sample maximum slightly. The formula for the median of the sample maximum (Eq. 86) provides a more accurate estimate of this sample extreme. Therefore, the following estimate for the mean of the attribute range will be used instead

$$\mathbb{E}(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m) \approx 2 \left[ \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m}\right)} - \Phi^{-1}\left(\frac{1}{m}\right) \right], \quad (91)$$

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

We have already determined the mean and variance (Eq. 30) for the  $L_q$  metric (Eq. 1) on standard normal data. Using the expected value of the sample maximum (Eq. 91), the variance of the sample maximum (Eq. 90), and the general formulas for the mean and variance of the max-min normalized distance distribution (Eq. 80), this leads us to the following asymptotic estimate for the distribution of the max-min normalized distances for standard normal data

$$D_{ij}^{(q*)} \sim \mathcal{N} \left( \frac{\mu_{D_{ij}}^{(q)}}{2\mu_{\max}^{(1)}(m)}, \frac{6\log(m)\sigma_{D_{ij}}^2}{\pi^2 + 24 \left[ \mu_{\max}^{(1)}(m) \right]^2 \log(m)} \right). \quad (92)$$

where  $m$  is the size of the sample from which the maximum is derived,  $\mu_{\max}^{(1)}$  is the median of the sample maximum (Eq. 86),  $\mu_{D_{ij}^{(q)}}$  is the expected  $L_q$  pairwise distance (Eq. 28), and  $\sigma_{D_{ij}^{(q)}}^2$  is the variance of the  $L_q$  pairwise distance (Eq. 29). The moments of the max-min normalized  $L_q$  distance metric in standard normal data (Eq. 92) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 2).

#### 2.4.2 Standard uniform data

Standard uniform data does not have an even density function. Due to the simplicity of the density function, however, we can derive the distribution of the maximum and minimum of a sample of size  $m$  explicitly. Using the general forms of the distribution functions of the maximum (Eq. 60) and minimum (Eq. 63), we have the following distribution functions for standard uniform data

$$F_{\max}(x) = x^m \quad (93)$$

and

$$F_{\min}(x) = 1 - (1 - x)^m, \quad (94)$$

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

Using the general forms of the density functions of the maximum (Eq. 61) and minimum (Eq. 64), we have the following density functions for standard uniform data

$$f_{\max}(x) = mx^{m-1} \quad (95)$$

and

$$f_{\min}(x) = m(1 - x)^{m-1}, \quad (96)$$

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

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

$$\begin{aligned} E(X_a^{\max}) &= \mu_{\max}^{(1)}(m) = \int_0^1 x f_{\max}(x) dx \\ &= \int_0^1 x [mx^{m-1}] dx \\ &= \frac{m}{m+1} \end{aligned} \quad (97)$$

and

$$\begin{aligned} E(X_a^{\min}) &= \mu_{\min}^{(1)}(m) = \int_0^1 x f_{\min}(x) dx \\ &= \int_0^1 x [m(1 - x)^{m-1}] dx \\ &= \frac{1}{m+1}, \end{aligned} \quad (98)$$

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

We can compute the second moment about the origin of the sample range as follows 618

$$\begin{aligned}
E[(X_a^{\max} - X_a^{\min})^2] &= E[(X_a^{\max})^2 - 2X_a^{\max}X_a^{\min} + (X_a^{\min})^2] \\
&= E[(X_a^{\max})^2] - 2E(X_a^{\max})E(X_a^{\min}) + E[(X_a^{\min})^2] \\
&= \mu_{\max}^{(2)}(m) - 2\mu_{\max}^{(1)}(m)\mu_{\min}^{(1)}(m) + \mu_{\min}^{(2)}(m) \\
&= \int_0^1 x^2 [mx^{m-1}]dx - 2\left(\frac{m}{m+1}\right)\left(\frac{1}{m+1}\right) \\
&\quad + \int_0^1 x^2 [m(1-x)^{m-1}]dx \\
&= \frac{m}{m+2} - \frac{2m}{(m+1)^2} + \frac{2}{(m+1)(m+2)} \\
&= \frac{m^3 - m + 2}{(m+2)(m+1)^2},
\end{aligned} \tag{99}$$

where  $m$  is the size of the sample from which extremes are derived. 619

Using the general asymptotic distribution of max-min normalized distances for any data type (Eq. 77) and the mean and variance (Eq. 40) of the standard  $L_q$  distance metric (Eq. 1), we have the following asymptotic estimate for the max-min normalized distance distribution for standard uniform data 620  
621  
622  
623

$$D_{ij}^{(q*)} \sim \mathcal{N}\left(\frac{(m+1)\mu_{D_{ij}^{(q)}}}{m-1}, \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(q)}}^2}{m^3 - m + 2}\right), \tag{100}$$

where  $m$  is the size of the sample from which extremes are derived,  $\mu_{D_{ij}^{(q)}}$  is the expected value (Eq. 38) of the  $L_q$  metric (Eq. 1) in standard uniform data, and  $\sigma_{D_{ij}^{(q)}}^2$  is the variance (Eq. 39) of the  $L_q$  metric (Eq. 1) in standard uniform data. The moments of the max-min normalized  $L_q$  distance metric in standard uniform data (Eq. 92) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 2). 624  
625  
626  
627  
628  
629

## 2.5 Range-Normalized Manhattan ( $q = 1$ ) 630

Using the general asymptotic results for mean and variance of max-min normalized distances in standard normal and standard uniform data (Eqs. 92 and 100) for any value of  $q \in \mathbb{N}$ , we can substitute a particular value of  $q$  in order to determine a more specified distribution for the normalized distance ( $D^{(q*)}$ , Eq. 58). The following results are for the max-min normalized Manhattan ( $q = 1$ ),  $D^{(1*)}$ , metric for both standard normal and standard uniform data. 631  
632  
633  
634  
635  
636

### 2.5.1 Standard normal data 637

Substituting  $q = 1$  into the asymptotic formula for the expected max-min normalized distance (Eq. 92), we derive the expected normalized Manhattan distance in standard normal data as follows 638  
639  
640

$$\begin{aligned}
E(D_{ij}^{(1*)}) &= \frac{\mu_{D_{ij}^{(1)}}}{2\mu_{\max}^{(1)}(m)} \\
&= \frac{p}{\sqrt{\pi}\mu_{\max}^{(1)}(m)},
\end{aligned} \tag{101}$$

where  $\mu_{\max}^{(1)}(m)$  is the expected attribute maximum (Eq. 86),  $m$  is the size of the sample from which the maximum is derived, and  $p$  is the total number of attributes. 641  
642

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

$$\begin{aligned}\text{Var}\left(D_{ij}^{(1*)}\right) &= \frac{6\log(m)\sigma_{D_{ij}^{(1)}}^2}{\pi^2 + 24\left[\mu_{\max}^{(1)}\right]^2\log(m)} \\ &= \frac{12p(\pi - 2)\log(m)}{\pi\left(\pi^2 + 24\left[\mu_{\max}^{(1)}\right]^2\log(m)\right)},\end{aligned}\tag{102}$$

where  $\mu_{\max}^{(1)}(m)$  is the expected attribute maximum (Eq. 86),  $m$  is the size of the sample from which the maximum is derived, and  $p$  is the total number of attributes. Similar to the variance of the standard Manhattan distance, the variance of the max-min normalized Manhattan distance is on the order of  $p$  for fixed instance dimension  $m$ . For fixed  $p$ , the variance (Eq. 102) vanishes as  $m$  grows without bound. If we fix  $m$ , the same variance increases monotonically with increasing  $p$ . The moments derived in this section (Eqs. 101 and 102) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 4).

### 2.5.2 Standard uniform data

Substituting  $q = 1$  into the asymptotic formula for the expected max-min pairwise distance (Eq. 100), we derive the expected normalized Manhattan distance in standard uniform data as

$$\begin{aligned}\mathbb{E}\left(D_{ij}^{(1*)}\right) &= \frac{(m+1)\mu_{D_{ij}^{(1)}}}{m-1} \\ &= \frac{(m+1)p}{3(m-1)},\end{aligned}\tag{103}$$

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

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

$$\begin{aligned}\text{Var}\left(D_{ij}^{(1*)}\right) &= \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(1)}}^2}{m^3 - m + 2} \\ &= \frac{(m+2)(m+1)^2p}{18(m^3 - m + 2)},\end{aligned}\tag{104}$$

where  $m$  is the size of the sample from which extremes are derived and  $p$  is the total number of attributes. Interestingly, the variance of the max-min normalized Manhattan distance in standard uniform data approaches  $p/18$  as  $m$  increases without bound for a fixed number of attributes  $p$ . This is the same asymptotic value to which the variance of the standard Manhattan distance (Eq. 43) converges. Therefore, large sample sizes make the variance of the normalized Manhattan distance approach the variance of the standard Manhattan distance in standard uniform data. The moments derived in this section (Eqs. 103 and 104) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 4).

## 2.6 Range-Normalized Euclidean ( $q = 2$ )

Analogous to the previous section, we use the asymptotic moment estimates for the max-min normalized metric ( $D^{(q*)}$ , Eq. 58) for standard normal (Eq. 92) and standard uniform (Eq. 100) data but specific to a range-normalized Euclidean metric ( $q = 2$ ).

### 2.6.1 Standard normal data

Substituting  $q = 2$  into the asymptotic formula for the expected max-min normalized pairwise distance (Eq. 92), we derive the expected normalized Euclidean distance in standard normal data as

$$\begin{aligned} \mathbb{E} \left( D_{ij}^{(2*)} \right) &= \frac{\mu_{D_{ij}^{(2)}}^{(1)}}{2\mu_{\max}^{(1)}(m)} \\ &= \frac{\sqrt{2p-1}}{2\mu_{\max}^{(1)}(m)}, \end{aligned} \quad (105)$$

where  $\mu_{\max}^{(1)}(m)$  is the expected attribute maximum (Eq. 86),  $m$  is the size of the sample from which the maximum is derived, and  $p$  is the total number of attributes.

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

$$\begin{aligned} \text{Var} \left( D_{ij}^{(2*)} \right) &= \frac{6\log(m)\sigma_{D_{ij}^{(2)}}^2}{\pi^2 + 24 \left[ \mu_{\max}^{(1)}(m) \right]^2 \log(m)} \\ &= \frac{6\log(m)}{\pi^2 + 24 \left[ \mu_{\max}^{(1)}(m) \right]^2 \log(m)}, \end{aligned} \quad (106)$$

where  $\mu_{\max}^{(1)}(m)$  is the expected attribute maximum (Eq. 86) and  $m$  is the size of the sample from which the maximum is derived. It is interesting to note that the variance (Eq. 106) vanishes as the sample size  $m$  increases without bound, which means that all distances will be tightly clustered about the mean (Eq. 105). This is different than the variance of the standard  $L_2$  metric (Eq. 52), which is asymptotically equal to 1. This could imply that any two pairwise distances computed with the max-min normalized Euclidean metric in a large sample space  $m$  may be indistinguishable, which is another curse of dimensionality. The moments derived in this section (Eqs. 105 and 106) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 4).

### 2.6.2 Standard uniform data

Substituting  $q = 2$  into the asymptotic formula for the expected max-min normalized pairwise distance (Eq. 100), we derive the expected normalized Euclidean distance in standard uniform data as

$$\begin{aligned} \mathbb{E} \left( D_{ij}^{(2*)} \right) &= \frac{(m+1)\mu_{D_{ij}^{(2)}}}{m-1} \\ &= \sqrt{\frac{p}{6} - \frac{7}{120} \left( \frac{m+1}{m-1} \right)}. \end{aligned} \quad (107)$$

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

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

$$\begin{aligned} \text{Var} \left( D_{ij}^{(2*)} \right) &= \frac{(m+2)(m+1)^2\sigma_{D_{ij}^{(2)}}^2}{m^3 - m + 2} \\ &= \frac{7(m+2)(m+1)^2}{120(m^3 - m + 2)}. \end{aligned} \quad (108)$$

where  $m$  is the size of the sample from which extremes are derived. Similar to the variance of max-min normalized Manhattan distances in standard uniform data (Eq. 104), the variance of normalized Euclidean distances approaches the variance of the standard Euclidean distances in uniform data (Eq. 55) as  $m$  increases without bound. That is, the variance of the max-min normalized Euclidean distance (Eq. 108) approaches  $7/120$  as  $m$  grows larger. The moments derived in this section (Eqs. 107 and 108) are summarized in a figure that is organized by metric, data type, statistic (mean or variance), and asymptotic formula (Fig. 4).

We summarize moment estimates in figures (Fig. 2-4) that contain all of our asymptotic results for both standard and max-min normalized  $L_q$  metrics in each data type we have considered. This includes our most general results for any combination of sample size  $m$ , number of attributes  $p$ , type of metric  $L_q$ , and data type (Fig. 2). From these more general derivations, we show the results of the standard  $L_1$  and  $L_2$  metrics for any combination of sample size  $m$ , number of attributes  $p$ , and data type (Fig. 3). Our last set of summarized results show asymptotics for the max-min normalized  $L_1$  and  $L_2$  metrics for any combination of sample size  $m$ , number of attributes  $p$ , and data type (Fig. 4). For both standard and max-min normalized  $L_2$  metrics (Fig. 3 and 4), the low-dimensional improved estimates of sample means (Eqs. 53 and 56) are used because they perform well at both low and high attribute dimension  $p$ .

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

**Fig 2. Summary of distance distribution derivations for standard normal ( $\mathcal{N}(0, 1)$ ) and standard uniform ( $\mathcal{U}(0, 1)$ ) data.** Asymptotic estimates are given for both standard (Eq. 1) and max-min normalized (Eq. 58)  $q$ -metrics. These estimates are relevant for all  $q \in \mathbb{N}$  and  $p \gg 1$  for which the normality assumption of distances holds.

**Fig 3. Asymptotic estimates of means and variances for the standard  $L_1$  and  $L_2$  ( $q = 1$  and  $q = 2$  in Fig. 2) distance distributions.** Estimates for both standard normal ( $\mathcal{N}(0, 1)$ ) and standard uniform ( $\mathcal{U}(0, 1)$ ) data are given.

**Fig 4. Asymptotic estimates of means and variances for the max-min normalized  $L_1$  and  $L_2$  distance distributions commonly used in Relief-based algorithms.** Estimates for both standard normal ( $\mathcal{N}(0, 1)$ ) and standard uniform ( $\mathcal{U}(0, 1)$ ) data are given. The cumulative distribution function of the standard normal distribution is represented by  $\Phi$ . Furthermore,  $\mu_{\max}^{(1)}(m)$  (Eq. 86) is the asymptotic median of the sample maximum from  $m$  standard normal random samples.

### 3 GWAS distance distributions

In genome-wide association studies (GWAS), data is encoded by the minor allele at a particular locus  $a$ , which is just the second most common allele (adenine-A, thymine-T, cytosine-C, or guanine-G) associated with a given attribute  $a$  in the data set. Attributes in GWAS data are single nucleotide polymorphisms (SNPs), or mutations involving the substitution, deletion, or insertion of one nucleotide at some point in the DNA sequence of an organism. These are common mutations that can affect how an individual reacts to certain pathogens or the susceptibility for certain diseases. Feature selection in GWAS is typically concerned with finding interacting SNPs that are associated with disease susceptibility [21].

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

$$X_{ia} = \begin{cases} 0 & \text{if there are no minor alleles at locus } a, \\ 1 & \text{if there is 1 minor allele at locus } a, \\ 2 & \text{if there are 2 minor alleles at locus } a. \end{cases} \quad (109)$$

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

$$d_{ij}^{\text{GM}}(a) = \begin{cases} 0 & \text{if } X_{ia} \neq X_{ja}, \\ 1 & \text{otherwise} \end{cases} \quad (110)$$

and

$$d_{ij}^{\text{AM}}(a) = \frac{1}{2} |X_{ia} - X_{ja}|. \quad (111)$$

More informative metrics may include differences at the nucleotide level for each allele by considering differences in the rates of transition and transversion mutations (Fig. 5). One such discrete metric that accounts for transitions (Ti) and transversions (Tv) was introduced in [7] and can be written as

$$d_{ij}^{\text{TiTv}}(a) = \begin{cases} 0 & \text{if } X_{ia} = X_{ja} \text{ and Ti/Tv,} \\ 1/4 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Ti,} \\ 1/2 & \text{if } |X_{ia} - X_{ja}| = 1 \text{ and Tv,} \\ 3/4 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Ti,} \\ 1 & \text{if } |X_{ia} - X_{ja}| = 2 \text{ and Tv.} \end{cases} \quad (112)$$

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

$$D_{ij}^{\text{GM}}(a) = \sum_{a \in \mathcal{A}} d_{ij}^{\text{GM}}(a), \quad (113)$$

$$D_{ij}^{\text{AM}}(a) = \sum_{a \in \mathcal{A}} d_{ij}^{\text{AM}}(a), \text{ or} \quad (114)$$

$$D_{ij}^{\text{TiTv}}(a) = \sum_{a \in \mathcal{A}} d_{ij}^{\text{TiTv}}(a). \quad (115)$$

Assuming that all data entries  $X_{ia}$  are independent and identically distributed, we have already shown that the distribution of pairwise distances is asymptotically normal regardless of data distribution and value of  $q$ . Therefore, the distance distributions induced by each of the GWAS metrics (Eqs. 110-112) are asymptotically normal. With this Gaussian limiting behavior, we will proceed by deriving the mean and variance for each distance distribution induced by these three GWAS metrics.

### 3.1 GM distance distribution

The simplest distance metric in nearest-neighbor feature selection in GWAS data is the genotype-mismatch (GM) distance metric (Eq. 113). The GM attribute diff (Eq. 110) indicates only whether two genotypes are the same or not. There are many ways two genotypes could differ, but this metric does not record this information. We will now derive the moments for the GM distance (Eq. 113), which are sufficient for defining its corresponding asymptotic distribution.

The expected value of the GM attribute diff metric (Eq. 110) is given by the following

$$\begin{aligned}
 E \left[ d_{ij}^{\text{GM}}(a) \right] &= \sum_{k=0}^1 k \cdot P \left[ d_{ij}^{\text{GM}}(a) = k \right] \\
 &= 0 \cdot P \left[ d_{ij}^{\text{GM}}(a) = 0 \right] + 1 \cdot P \left[ d_{ij}^{\text{GM}}(a) = 1 \right] \\
 &= P \left[ d_{ij}^{\text{GM}}(a) = 1 \right] \\
 &= 2P[X_{ia} = 0, X_{ja} = 1] + 2P[X_{ia} = 1, X_{ja} = 2] + 2P[X_{ia} = 0, X_{ja} = 2] \\
 &= 4(1 - f_a)^3 f_a + 4(1 - f_a) f_a^3 + 2(1 - f_a)^2 f_a^2 \\
 &= 2 \left[ 2(1 - f_a)^3 f_a + 2(1 - f_a) f_a^3 + (1 - f_a)^2 f_a^2 \right] \\
 &= 2F^{\text{GM}}(a),
 \end{aligned} \tag{116}$$

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

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

$$\begin{aligned}
 E(D_{ij}^{\text{GM}}) &= E \left( \sum_{a \in \mathcal{A}} d_{ij}^{\text{GM}}(a) \right) \\
 &= \sum_{a \in \mathcal{A}} E \left[ d_{ij}^{\text{GM}}(a) \right] \\
 &= 2 \sum_{a \in \mathcal{A}} F^{\text{GM}}(a),
 \end{aligned} \tag{117}$$

where  $F^{\text{GM}}(a) = 2(1 - f_a)^3 f_a + 2(1 - f_a) f_a^3 + (1 - f_a)^2 f_a^2$  and  $f_a$  is the probability of a minor allele occurring at locus  $a$ . We see that the expected GM pairwise distance (Eq. 117) relies only on the minor allele probabilities  $f_a$  for all  $a \in \mathcal{A}$ . In real data, we can easily determine these probabilities by dividing the total number of minor alleles at locus  $a$  by the twice the number of instances  $m$ . To be more explicit, this is just

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

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



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

$$\begin{aligned}
\mathbb{E} \left[ (D_{ij}^{\text{GM}})^2 \right] &= \mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} d_{ij}^{\text{GM}}(a) \right)^2 \right] \\
&= \mathbb{E} \left[ \sum_{a \in \mathcal{A}} \left( d_{ij}^{\text{GM}}(a) \right)^2 \right] + 2\mathbb{E} \left[ \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} d_{ij}^{\text{GM}}(r) \cdot d_{ij}^{\text{GM}}(s) \right] \\
&= \sum_{a \in \mathcal{A}} \left( \sum_{k=0}^1 k^2 \cdot \mathbb{P} \left[ d_{ij}^{\text{GM}}(a) = k \right] \right) \\
&\quad + 2 \sum_{a \in \mathcal{A}} \sum_{s \leq r-1} \left( \sum_{k=0}^1 k \cdot \mathbb{P} \left[ d_{ij}^{\text{GM}}(r) = k \right] \right) \cdot \left( \sum_{k=0}^1 k \cdot \mathbb{P} \left[ d_{ij}^{\text{GM}}(s) = k \right] \right) \\
&= 2 \sum_{a \in \mathcal{A}} F^{\text{GM}}(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} F^{\text{GM}}(\lambda),
\end{aligned} \tag{118}$$

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

Using the first (Eq. 117) and second (Eq. 118) raw moments of the GM distance, the variance is given by

$$\begin{aligned}
\text{Var} (D_{ij}^{\text{GM}}) &= \mathbb{E} \left[ (D_{ij}^{\text{GM}})^2 \right] - [\mathbb{E} (D_{ij}^{\text{GM}})]^2 \\
&= 2 \sum_{a \in \mathcal{A}} F^{\text{GM}}(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} F^{\text{GM}}(\lambda) - 4 \left( \sum_{a \in \mathcal{A}} F^{\text{GM}}(a) \right)^2 \\
&= 2 \sum_{a \in \mathcal{A}} F^{\text{GM}}(a) - 4 \sum_{a \in \mathcal{A}} [F^{\text{GM}}(a)]^2 \\
&= 2 \sum_{a \in \mathcal{A}} F^{\text{GM}}(a)[1 - 2F^{\text{GM}}(a)],
\end{aligned} \tag{119}$$

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

With the mean and variance estimates (Eqs. 117 and 119), the asymptotic GM distance distribution is given by the following

$$D_{ij}^{\text{GM}} \sim \mathcal{N} \left( 2 \sum_{a \in \mathcal{A}} F^{\text{GM}}(a), 2 \sum_{a \in \mathcal{A}} F^{\text{GM}}(a)[1 - 2F^{\text{GM}}(a)] \right), \tag{120}$$

where  $F^{\text{GM}}(a) = 2(1 - f_a)^3 f_a + 2(1 - f_a)f_a^3 + (1 - f_a)^2 f_a^2$  and  $f_a$  is the probability of a minor allele occurring at locus  $a$ . This GM distribution holds for random independent GWAS data with minor allele probabilities  $f_a$  and binomial samples  $X_{ia} \sim \mathcal{B}(2, f_a)$  for all  $a \in \mathcal{A}$ . Next we consider the distance distribution for an AM metric, which incorporates differences at the allele level and contains more information than genotype differences.

### 3.2 AM distance distribution

As we have mentioned previously, the AM attribute diff metric (Eq. 111) is slightly more dynamic than the GM metric because the AM metric accounts for differences between

the alleles of two genotypes. In this section, we derive moments of the AM distance metric (Eq. 114) that adequately define its corresponding asymptotic distribution.

The expected value of the AM attribute diff metric (Eq. 111) is given by the following

$$\begin{aligned}
\mathbb{E} \left[ d_{ij}^{\text{AM}}(a) \right] &= \sum_{k \in \mathcal{D}} k \cdot \mathbb{P} \left[ d_{ij}^{\text{AM}}(a) = k \right] \\
&= 0 \cdot \mathbb{P} \left[ d_{ij}^{\text{AM}}(a) = 0 \right] + \frac{1}{2} \cdot \mathbb{P} \left[ d_{ij}^{\text{AM}}(a) = \frac{1}{2} \right] + 1 \cdot \mathbb{P} \left[ d_{ij}^{\text{AM}}(a) = 1 \right] \\
&= \frac{1}{2} (2\mathbb{P} [X_{ia} = 0, X_{ja} = 1] + 2\mathbb{P} [X_{ia} = 1, X_{ja} = 2]) \\
&\quad + 2\mathbb{P} [X_{ia} = 0, X_{ja} = 2] \\
&= \mathbb{P} [X_{ia} = 0, X_{ja} = 1] + \mathbb{P} [X_{ia} = 1, X_{ja} = 2] + 2\mathbb{P} [X_{ia} = 0, X_{ja} = 2] \\
&= 2(1 - f_a)^3 f_a + 2(1 - f_a) f_a^3 + 2(1 - f_a)^2 f_a^2 \\
&= 2 \left[ (1 - f_a)^3 f_a + (1 - f_a) f_a^3 + (1 - f_a)^2 f_a^2 \right] \\
&= 2F^{\text{AM}}(a),
\end{aligned} \tag{121}$$

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

Using the expected AM attribute diff (Eq. 121), the expected pairwise AM distance (Eq. 114) between instances  $i, j \in \mathcal{I}$  is given by

$$\begin{aligned}
\mathbb{E} (D_{ij}^{\text{AM}}) &= \mathbb{E} \left( \sum_{a \in \mathcal{A}} d_{ij}^{\text{AM}}(a) \right) \\
&= \sum_{a \in \mathcal{A}} \mathbb{E} \left[ d_{ij}^{\text{AM}}(a) \right] \\
&= 2 \sum_{a \in \mathcal{A}} F^{\text{AM}}(a).
\end{aligned} \tag{122}$$

where  $F^{\text{AM}}(a) = (1 - f_a)^3 f_a + (1 - f_a) f_a^3 + (1 - f_a)^2 f_a^2$  and  $f_a$  is the probability of a minor allele occurring at locus  $a$ . Similar to GM distances, the expected AM distance (Eq. 122) depends only on the minor allele probabilities  $f_a$  for all  $a \in \mathcal{A}$ . This is to be expected because, although the AM metric is more informative, it still only accounts for simple differences between nucleotides of two instances  $i, j \in \mathcal{I}$  at some locus  $a$ .

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

$$\begin{aligned}
\mathbb{E} \left[ (D_{ij}^{\text{AM}})^2 \right] &= \mathbb{E} \left[ \left( \sum_{a \in \mathcal{A}} d_{ij}^{\text{AM}}(a) \right)^2 \right] \\
&= \mathbb{E} \left[ \sum_{a \in \mathcal{A}} \left( d_{ij}^{\text{AM}}(a) \right)^2 \right] + 2\mathbb{E} \left[ \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} d_{ij}^{\text{AM}}(r) \cdot d_{ij}^{\text{AM}}(s) \right] \\
&= \sum_{a \in \mathcal{A}} \left( \sum_{k \in \mathcal{D}} k^2 \cdot \mathbb{P} \left[ d_{ij}^{\text{AM}}(a) = k \right] \right) \\
&\quad + 2 \sum_{a \in \mathcal{A}} \sum_{s \leq r-1} \left( \sum_{k \in \mathcal{D}} k \cdot \mathbb{P} \left[ d_{ij}^{\text{AM}}(r) = k \right] \right) \cdot \left( \sum_{k \in \mathcal{D}} k \cdot \mathbb{P} \left[ d_{ij}^{\text{AM}}(s) = k \right] \right) \\
&= \sum_{a \in \mathcal{A}} G^{\text{AM}}(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} F^{\text{AM}}(\lambda),
\end{aligned} \tag{123}$$

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

Using the first (Eq. 122) and second (Eq. 123) raw moments of the asymptotic AM  
distance distribution, the variance is given by

$$\begin{aligned} \text{Var}(D_{ij}^{\text{AM}}) &= \mathbb{E}[(D_{ij}^{\text{AM}})^2] - [\mathbb{E}(D_{ij}^{\text{AM}})]^2 \\ &= \sum_{a \in \mathcal{A}} G^{\text{AM}}(a) + 8 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} F^{\text{AM}}(\lambda) - 4 \left( \sum_{a \in \mathcal{A}} F^{\text{AM}}(a) \right)^2 \\ &= \sum_{a \in \mathcal{A}} G^{\text{AM}}(a) - 4 \sum_{a \in \mathcal{A}} [F^{\text{AM}}(a)]^2 \\ &= \sum_{a \in \mathcal{A}} \left( G^{\text{AM}}(a) - 4 [F^{\text{AM}}(a)]^2 \right), \end{aligned} \quad (124)$$

where  $G^{\text{AM}}(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a) + 2(1 - f_a)^2 f_a^2$ ,  
 $F^{\text{AM}}(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a) + (1 - f_a)^2 f_a^2$ , and  $f_a$  is the probability of a minor  
allele occurring at locus  $a$ . Similar to the mean (Eq. 122), the variance just depends on  
minor allele probabilities  $f_a$  for all  $a \in \mathcal{A}$ .

With the mean (Eq. 122) and variance (Eq. 124) estimates of AM distances, the  
asymptotic AM distance distribution is given by the following

$$D_{ij}^{\text{AM}} \sim \mathcal{N} \left( 2 \sum_{a \in \mathcal{A}} F^{\text{AM}}(a), \sum_{a \in \mathcal{A}} \left( G^{\text{AM}}(a) - 4 [F^{\text{AM}}(a)]^2 \right) \right), \quad (125)$$

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

This concludes our analysis of the AM metric in GWAS data when the independence  
assumption holds for minor allele probabilities  $f_a$  and binomial samples  $\mathcal{B}(2, f_a)$  for all  
 $a \in \mathcal{A}$ . In the next section, we derive more complex asymptotic results for the TiTv  
distance metric (Eq. 115).

### 3.3 TiTv distance distribution

The TiTv metric allows for one to account for both genotype mismatch, allele mismatch,  
transition, and transversion. However, this added dimension of information requires  
knowledge of the nucleotide makeup at a particular locus. A sufficient condition to  
compute the TiTv metric between instances  $i, j \in \mathcal{I}$  is that we know whether the  
nucleotides associated with a particular locus  $a$  are both purines (PuPu), purine and  
pyrimidine (PuPy), or both pyrimidines (PyPy). We illustrate all possibilities for  
transitions and transversions in a diagram (Fig. 5). Purines (A and G) and pyrimidines  
(C and T) are shown at the top and bottom, respectively. Transitions occur in the cases  
of PuPu and PyPy, while transversion occurs only with PuPy encoding.

**Fig 5. 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.

837

This additional encoding is always given in a particular GWAS data set, which leads us to consider the probabilities of PuPu, PuPy, and PyPy. These will be necessary to determine asymptotics for the TiTv distance metric. Let  $\gamma_0$ ,  $\gamma_1$ , and  $\gamma_2$  denote the probabilities of PuPu, PuPy, and PyPy, respectively, for the  $p$  loci of data matrix  $X$ . In real data, there are approximately twice as many transitions as there are transversions. That is, the probability of a transition  $P(\text{Ti})$  is approximately twice the probability of transversion  $P(\text{Tv})$ . It is likely that any particular data set will not satisfy this criterion exactly. In this general case, we have  $P(\text{Ti})$  being equal to some multiple  $\eta$  times  $P(\text{Tv})$ . In order to enforce this general constraint in simulated data, we define the following set of equalities

$$\gamma_0 + \gamma_1 + \gamma_2 = 1, \quad (126)$$

$$P(\text{Ti}) - \eta P(\text{Tv}) = 0. \quad (127)$$

The sum-to-one constraint (Eq. 126) is natural in this context because there are only three possible genotype encodings at a particular locus, which are PuPu, PuPy, and PyPy. Solving the Ti/Tv ratio constraint (Eq. 127) for  $\eta$  gives

838

839

840

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

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

841

842

843

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

844

845

$$P(\text{Tv}) = \gamma_1. \quad (128)$$

Using the sum-to-one constraint (Eqs. 126) and the probability of transversion (Eq. 127), the probability of a transition occurring at locus  $a$  is computed as follows

846

847

$$P(\text{Ti}) = \gamma_0 + \gamma_2. \quad (129)$$

Also using the sum-to-one constraint (Eq. 126) and the Ti/Tv ratio constraint (Eq. 127), 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

848

849

850

$$\gamma_0 \sim \mathcal{U}\left(\varepsilon, \frac{\eta}{\eta+1} - \varepsilon\right), \quad (130)$$

where  $\varepsilon$  is some small positive real number. 851

Then it immediately follows that we have 852

$$\gamma_2 = \frac{\eta}{\eta + 1} - \gamma_0. \quad (131)$$

However, we can derive the mean and variance of the distance distribution induced by the TiTv metric without specifying any relationship between  $\gamma_0$ ,  $\gamma_1$ , and  $\gamma_2$ . We proceed by computing  $P[d_{ij}^{\text{TiTv}}(a) = k]$  for each  $k \in \mathcal{D} = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ . Let  $y$  represent a random sample of size  $p$  from  $\{0, 1, 2\}$ , where 853  
854  
855  
856

$$y_a = \begin{cases} 0 & \text{if locus } a \text{ is PuPu,} \\ 1 & \text{if locus } a \text{ is PuPy,} \\ 2 & \text{if locus } a \text{ is PyPy.} \end{cases} \quad (132)$$

We derive  $P[d_{ij}^{\text{TiTv}}(a) = 0]$  as follows 857

$$\begin{aligned} P[d_{ij}^{\text{TiTv}}(a) = 0] &= P[y_a = 0, X_{ia} = X_{ja}] \\ &\quad + P[y_a = 1, X_{ia} = X_{ja}] \\ &\quad + P[y_a = 2, X_{ia} = X_{ja}] \\ &= \gamma_0 [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &\quad + \gamma_1 [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &\quad + \gamma_2 [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &= (\gamma_0 + \gamma_1 + \gamma_2) [(1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2] \\ &= (1 - f_a)^2 + 4f_a(1 - f_a) + f_a^2, \end{aligned} \quad (133)$$

where  $f_a$  is the probability of a minor allele occurring at locus  $a$ . 858

We derive  $P[d_{ij}^{\text{TiTv}}(a) = \frac{1}{4}]$  as follows 859

$$\begin{aligned} P[d_{ij}^{\text{TiTv}}(a) = \frac{1}{4}] &= 2P[y_a = 0, X_{ia} = 0, X_{ja} = 1] \\ &\quad + 2P[y_a = 0, X_{ia} = 1, X_{ja} = 2] \\ &\quad + 2P[y_a = 2, X_{ia} = 0, X_{ja} = 1] \\ &\quad + 2P[y_a = 2, X_{ia} = 1, X_{ja} = 2] \\ &= 4\gamma_0(1 - f_a)^3 f_a + 4\gamma_0 f_a^3(1 - f_a) + 4\gamma_2(1 - f_a)^3 f_a \\ &\quad + 4\gamma_2 f_a^3(1 - f_a) \\ &= 4\gamma_0 [(1 - f_a)^3 f_a + f_a^3(1 - f_a)] \\ &\quad + 4\gamma_2 [(1 - f_a)^3 f_a + f_a^3(1 - f_a)] \\ &= 4(\gamma_0 + \gamma_2) [(1 - f_a)^3 f_a + f_a^3(1 - f_a)], \end{aligned} \quad (134)$$

where  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu occurring at any locus  $a$ , and  $\gamma_2$  is the probability of PyPy occurring at any locus  $a$ . 860  
861  
862

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

$$\begin{aligned} P \left[ d_{ij}^{\text{TiTv}}(a) = \frac{1}{2} \right] &= 2P[y_a = 1, X_{ia} = 0, X_{ja} = 1] \\ &\quad + 2P[y_a = 1, X_{ia} = 1, X_{ja} = 2] \\ &= 4\gamma_1(1 - f_a)^3 f_a + 4\gamma_1 f_a^3(1 - f_a) \\ &= 4\gamma_1 [(1 - f_a)^3 f_a + f_a^3(1 - f_a)], \end{aligned} \quad (135)$$

where  $f_a$  is the probability of a minor allele occurring at locus  $a$  and  $\gamma_1$  is the probability of PuPy occurring at any locus  $a$ .

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

$$\begin{aligned} P \left[ d_{ij}^{\text{TiTv}}(a) = \frac{3}{4} \right] &= 2P[y_a = 0, X_{ia} = 0, X_{ja} = 2] \\ &\quad + 2P[y_a = 2, X_{ia} = 0, X_{ja} = 2] \\ &= 2\gamma_0(1 - f_a)^2 f_a^2 + 2\gamma_2(1 - f_a)^2 f_a^2 \\ &= 2(\gamma_0 + \gamma_2)(1 - f_a)^2 f_a^2, \end{aligned} \quad (136)$$

where  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu occurring at any locus  $a$ , and  $\gamma_2$  is the probability of PyPy occurring at any locus  $a$ .

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

$$\begin{aligned} P \left[ d_{ij}^{\text{TiTv}}(a) = 1 \right] &= 2P[y_a = 1, X_{ia} = 0, X_{ja} = 2] \\ &= 2\gamma_1(1 - f_a)^2 f_a^2, \end{aligned} \quad (137)$$

where  $f_a$  is the probability of a minor allele occurring at locus  $a$  and  $\gamma_1$  is the probability of PuPy occurring at any locus  $a$ .

Using the TiTv diff probabilities (Eqs. 133-137), we compute the expected TiTv distance between instances  $i, j \in \mathcal{I}$  as follows

$$\begin{aligned} E(D_{ij}^{\text{TiTv}}) &= \sum_{a \in \mathcal{A}} \left( \sum_{k \in \mathcal{D}} k \cdot P \left[ d_{ij}^{\text{TiTv}}(a) = k \right] \right) \\ &= (\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}} [(1 - f_a)^3 f_a + f_a^3(1 - f_a)] \\ &\quad + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} (1 - f_a)^2 f_a^2 \\ &= (\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}} F^{\text{TiTv}}(a) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G^{\text{TiTv}}(a), \end{aligned} \quad (138)$$

where  $F^{\text{TiTv}}(a) = (1 - f_a)^3 f_a + f_a^3(1 - f_a)$ ,  $G^{\text{TiTv}}(a) = (1 - f_a)^2 f_a^2$ ,  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu occurring at any locus  $a$ ,  $\gamma_1$  is the probability of PuPy occurring at any locus  $a$ , and  $\gamma_2$  is the probability of PyPy occurring at any locus  $a$ . In contrast to the expected GM and AM distances (Eqs. 117 and 122), the expected TiTv distance (Eq. 138) depends on minor allele probabilities  $f_a$  for all  $a \in \mathcal{A}$  and the genotype encoding probabilities  $\gamma_0, \gamma_1$ , and  $\gamma_2$ .

$$\begin{aligned}
E \left[ (D_{ij}^{\text{TiTv}})^2 \right] &= E \left[ \left( \sum_{a \in \mathcal{A}} d_{ij}^{\text{TiTv}}(a) \right)^2 \right] \\
&= E \left[ \sum_{a \in \mathcal{A}} \left( d_{ij}^{\text{TiTv}}(a) \right)^2 \right] + 2E \left[ \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} d_{ij}^{\text{TiTv}}(r) \cdot d_{ij}^{\text{TiTv}}(s) \right] \\
&= \sum_{a \in \mathcal{A}} \left( \sum_{k \in \mathcal{D}} k^2 \cdot P \left[ d_{ij}^{\text{TiTv}}(a) = k \right] \right) \\
&\quad + 2 \sum_{a \in \mathcal{A}} \sum_{s \leq r-1} \left( \sum_{k \in \mathcal{D}} k \cdot P \left[ d_{ij}^{\text{TiTv}}(r) = k \right] \right) \cdot \left( \sum_{k \in \mathcal{D}} k \cdot P \left[ d_{ij}^{\text{TiTv}}(s) = k \right] \right) \\
&= \left[ \frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1 \right] \sum_{a \in \mathcal{A}} F^{\text{TiTv}}(a) + \left[ \frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G^{\text{TiTv}}(a) \\
&\quad + 2 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} \left( [\gamma_0 + \gamma_2 + 2\gamma_1] F^{\text{TiTv}}(\lambda) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] G^{\text{TiTv}}(\lambda) \right),
\end{aligned} \tag{139}$$

where  $F^{\text{TiTv}}(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a)$ ,  $G^{\text{TiTv}}(a) = (1 - f_a)^2 f_a^2$ ,  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu occurring at any locus  $a$ ,  $\gamma_1$  is the probability of PuPy occurring at any locus  $a$ , and  $\gamma_2$  is the probability of PyPy occurring at any locus  $a$ .

Using the first (Eq. 138) and second (Eq. 139) raw moments of the TiTv distance, the variance is given by

$$\begin{aligned}
\text{Var} (D_{ij}^{\text{TiTv}}) &= E \left[ (D_{ij}^{\text{TiTv}})^2 \right] - [E (D_{ij}^{\text{TiTv}})]^2 \\
&= \left[ \frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1 \right] \sum_{a \in \mathcal{A}} F^{\text{TiTv}}(a) + \left[ \frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G^{\text{TiTv}}(a) \\
&\quad + 2 \sum_{r \in \mathcal{A}} \sum_{s \leq r-1} \prod_{\lambda \in \{r, s\}} \left( [\gamma_0 + \gamma_2 + 2\gamma_1] F^{\text{TiTv}}(\lambda) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] G^{\text{TiTv}}(\lambda) \right) \\
&\quad - \left( [\gamma_0 + \gamma_2 + 2\gamma_1] \sum_{a \in \mathcal{A}} F^{\text{TiTv}}(a) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G^{\text{TiTv}}(a) \right)^2 \\
&= \left[ \frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1 \right] \sum_{a \in \mathcal{A}} F^{\text{TiTv}}(a) + \left[ \frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G^{\text{TiTv}}(a) \\
&\quad - \sum_{a \in \mathcal{A}} \left( [\gamma_0 + \gamma_2 + 2\gamma_1] F^{\text{TiTv}}(a) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] G^{\text{TiTv}}(a) \right)^2,
\end{aligned} \tag{140}$$

where  $F^{\text{TiTv}}(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a)$ ,  $G^{\text{TiTv}}(a) = (1 - f_a)^2 f_a^2$ ,  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu occurring at any locus  $a$ ,  $\gamma_1$  is the probability of PuPy occurring at any locus  $a$ , and  $\gamma_2$  is the probability of PyPy occurring at any locus  $a$ .

With the mean (Eq. 138) and variance (Eq. 140) estimates, the asymptotic TiTv

distance distribution is given by the following

$$D_{ij}^{\text{TiTv}} \sim \mathcal{N} \left( (\gamma_0 + \gamma_2 + 2\gamma_1) \sum_{a \in \mathcal{A}} F^{\text{TiTv}}(a) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G^{\text{TiTv}}(a), \right. \\ \left. \left[ \frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1 \right] \sum_{a \in \mathcal{A}} F^{\text{TiTv}}(a) + \left[ \frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] \sum_{a \in \mathcal{A}} G^{\text{TiTv}}(a) \right. \\ \left. - \sum_{a \in \mathcal{A}} \left( [\gamma_0 + \gamma_2 + 2\gamma_1] F^{\text{TiTv}}(a) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] G^{\text{TiTv}}(a) \right)^2 \right), \quad (141)$$

where  $F^{\text{TiTv}}(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a)$ ,  $G^{\text{TiTv}}(a) = (1 - f_a)^2 f_a^2$ ,  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu occurring at any locus  $a$ ,  $\gamma_1$  is the probability of PuPy occurring at any locus  $a$ , and  $\gamma_2$  is the probability of PyPy occurring at any locus  $a$ .

Given upper and lower bounds  $l$  and  $u$ , respectively, of the success probability sampling interval, the average success probability (or average MAF) is computed as follows

$$\bar{f}_a = \frac{1}{2}(l + u). \quad (142)$$

The maximum TiTv distance occurs at  $\bar{f}_a = 0.5$  for any fixed Ti/Tv ratio  $\eta$  (Eq. 127), which is the inflection point about which the minor allele changes at locus  $a$  (Fig. 6). If few minor alleles are present ( $\bar{f}_a \rightarrow 0$ ), the predicted TiTv distance approaches 0. The same is true after the minor allele switches ( $\bar{f}_a \rightarrow 1$ ). To explore how TiTv distance changes with increased minor allele frequency, we fixed the Ti/Tv ratio  $\eta$  and generated simulated TiTv distances for  $\bar{f}_a = 0.055, 0.150, 0.250$ , and  $0.350$  (Fig. 7A). For fixed  $\eta$ , TiTv distance increases significantly with increased  $\bar{f}_a$ . We similarly fixed the average minor allele frequency  $\bar{f}_a$  and generated simulated TiTv distances for  $\eta = \text{Ti/Tv} = 0.5, 1, 1.5$ , and  $2$  (Fig. 7C). The TiTv distance decreases slightly with increased  $\eta = \text{Ti/Tv}$ . As  $\eta \rightarrow 0^+$ , 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  $\eta \rightarrow 2^-$  because the data is approaching approximately twice as many Ti as there are Tv, which is typical for GWAS data in humans.

**Fig 6. Predicted average TiTv distance as a function of average minor allele frequency  $\bar{f}_a$  (see Eq. 142).** Success probabilities  $f_a$  are drawn from a sliding window interval from 0.01 to 0.9 in increments of about 0.009 and  $m = p = 100$ . For  $\eta = 0.1$ , where  $\eta$  is the Ti/Tv ratio given by Eq. 126, Tv is ten times more likely than Ti and results in larger distance. Increasing to  $\eta = 1$ , Tv and Ti are equally likely and the distance is lower. In line with real data for  $\eta = 2$ , Tv is half as likely as Ti so the distances are relatively small.



We also compared theoretical and sample moments as a function of  $\eta = \text{Ti/Tv}$  and  $\bar{f}_a$  for the TiTv distance metric (Fig. 7B and D). We fixed  $\bar{f}_a$  and computed the theoretical and simulated moments as a function of  $\eta$  (Fig. 7B). Theoretical average TiTv distance, given by Eq. 138, and simulated TiTv average distance are approximately equal as  $\eta$  increases. Theoretical standard deviation, given by Eq. 140, and simulated TiTv standard deviation differ slightly. We also fixed  $\eta$  and computed theoretical and sample moments as a function of  $\bar{f}_a$  (Fig. 7D). In this case, there is approximate agreement with simulated and theoretical moments as  $\bar{f}_a$  increases.

**Fig 7. Density curves and moments of TiTv distance as a function of average MAF  $\bar{f}_a$ , given by Eq. 142, and Ti/Tv ratio  $\eta$ , given by Eq. 127.** We fix  $m = p = 100$  for all simulated TiTv distances. **(A)** For fixed  $\bar{f}_a = 0.055$ , TiTv distance density is plotted as a function of increasing  $\eta$ . TiTv distance decreases as  $\eta$  increases. For  $\eta = \text{Ti/Tv} = 0.5$ , there are twice as many transversions as there are transitions. On the other hand,  $\eta = \text{Ti/Tv} = 2$  indicates that there are half as many transversions as transitions. Since transversions encode a larger magnitude distance than transitions, this behavior is expected. **(B)** Simulated and predicted mean  $\pm$  SD are shown as a function of increasing Ti/Tv ratio  $\eta$ . Distance decreases as Ti/Tv increases. Theoretical and simulated moments are approximately the same. **(C)** For fixed  $\eta = 2$ , TiTv distance density is plotted as a function of increasing  $\bar{f}_a$ . TiTv distance increases as  $\bar{f}_a$  approaches maximum of 0.5, which means that there is about the same frequency of minor alleles as major alleles. **(D)** Simulated and predicted mean  $\pm$  SD as a function of increasing average MAF  $\bar{f}_a$ . Distance increases as the number of minor alleles increases. Theoretical and simulated moments are approximately the same.

We summarize our moment estimates for GWAS distance metrics (Eqs. 113-115) (Fig. 8) organized by metric, statistic (mean or variance), and asymptotic formula. Next we consider the important case of distributions of GWAS distances projected onto a single attribute (Eqs. 110-112).

**Fig 8. Asymptotic estimates of means and variances of genotype mismatch (GM) (Eq. 113), allele mismatch (AM) (Eq. 114), and transition-transversion (TiTv) (Eq. 115) distance metrics in GWAS data** ( $p \gg 1$ ). GWAS data  $X_{ia} \sim \mathcal{B}(2, f_a)$ , where  $f_a$  for all  $a \in \mathcal{A}$  are the probabilities of a minor allele occurring at locus  $a$ . For the TiTv distance metric, we have the additional encoding that uses  $\gamma_0 = \text{P}(\text{PuPu})$ ,  $\gamma_1 = \text{P}(\text{PuPy})$ , and  $\gamma_2 = \text{P}(\text{PyPy})$ .

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

We previously derived the exact distribution of the one-dimensional projected distance onto an attribute in continuous data (Section 2.2.3), which is used as the predictor in NPDR to calculate relative attribute importance in the form of standardized beta coefficients. GWAS data and the metrics we have considered are discrete. Therefore, we derive the density function for each diff metric (Eqs. 110-112), which also serves as the probability distribution for each metric, respectively.

The support of the GM metric (Eq. 110) is simply  $\{0, 1\}$ , so we derive the probability,  $\text{P} \left[ d_{ij}^{\text{GM}}(a) = k \right]$ , of this diff taking on each of these two possible values.

First, the probability that the GM diff is equal to zero is given by

$$\begin{aligned} f_{\text{GM}}(0; f_a) &= \text{P} \left[ d_{ij}^{\text{GM}}(a) = 0 \right] = \text{P} (X_{ia} = 0, X_{ja} = 0) + \text{P} (X_{ia} = 1, X_{ja} = 1) \\ &\quad + \text{P} (X_{ia} = 2, X_{ja} = 2) \quad (143) \\ &= (1 - f_a)^4 + 4f_a^2 (1 - f_a)^2 + f_a^4, \end{aligned}$$

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

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

$$\begin{aligned} f_{\text{GM}}(1; f_a) &= \text{P} \left[ d_{ij}^{\text{GM}}(a) = 1 \right] = 2\text{P} (X_{ia} = 0, X_{ja} = 1) + 2\text{P} (X_{ia} = 1, X_{ja} = 2) \\ &\quad + 2\text{P} (X_{ia} = 0, X_{ja} = 2) \quad (144) \\ &= 4(1 - f_a)^3 f_a + 4f_a^3 (1 - f_a) + 2f_a^2 (1 - f_a)^2, \end{aligned}$$

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

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

$$f_{\text{GM}}(d; f_a) = \begin{cases} (1 - f_a)^4 + 4f_a^2 (1 - f_a)^2 + f_a^4 & d = 0, \\ 4(1 - f_a)^3 f_a + 4f_a^3 (1 - f_a) + 2f_a^2 (1 - f_a)^2 & d = 1, \end{cases} \quad (145)$$

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

The mean and variance of this GM diff distribution can easily be derived using this newly determined density function (Eq. 145). The average GM diff is given by the following

$$\text{E} \left[ d_{ij}^{\text{GM}}(a) \right] = 2F^{\text{GM}}(a), \quad (146)$$

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

The variance of the GM diff metric is given by

$$\text{Var} \left[ d_{ij}^{\text{GM}}(a) \right] = 2F^{\text{GM}}(a) [1 - 2F^{\text{GM}}(a)], \quad (147)$$

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

The support of the AM metric (Eq. 111) is  $\{0, 1/2, 1\}$ . Beginning with the probability of the AM diff being equal to 0, we have the following probability

$$\begin{aligned} f_{\text{AM}}(0; f_a) &= \text{P} \left[ d_{ij}^{\text{AM}}(a) = 0 \right] = \text{P} (X_{ia} = 0, X_{ja} = 0) + \text{P} (X_{ia} = 1, X_{ja} = 1) \\ &\quad + \text{P} (X_{ia} = 2, X_{ja} = 2) \quad (148) \\ &= (1 - f_a)^4 + 4f_a^2 (1 - f_a)^2 + f_a^4, \end{aligned}$$

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

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

$$\begin{aligned} f_{\text{AM}}(1/2; f_a) &= \text{P} \left[ d_{ij}^{\text{AM}}(a) = 1/2 \right] = 2\text{P} (X_{ia} = 0, X_{ja} = 1) + 2\text{P} (X_{ia} = 1, X_{ja} = 2) \\ &= 4(1 - f_a)^3 f_a + 4f_a^3 (1 - f_a), \end{aligned} \quad (149)$$

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

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

$$f_{\text{AM}}(1; f_a) = \text{P} \left[ d_{ij}^{\text{AM}}(a) = 1 \right] = 2\text{P} (X_{ia} = 0, X_{ja} = 2) = 2f_a^2 (1 - f_a)^2, \quad (150)$$

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

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

$$f_{\text{AM}}(d; f_a) = \begin{cases} (1 - f_a)^4 + 4f_a^2 (1 - f_a)^2 + f_a^4 & d = 0, \\ 4(1 - f_a)^3 f_a + 4f_a^3 (1 - f_a) & d = 1/2, \\ 2f_a^2 (1 - f_a)^2 & d = 1, \end{cases} \quad (151)$$

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

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

$$\text{E} \left[ d_{ij}^{\text{AM}}(a) \right] = 2F^{\text{AM}}(a), \quad (152)$$

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

The variance of the AM diff metric is given by

$$\text{Var} \left[ d_{ij}^{\text{AM}}(a) \right] = G^{\text{AM}}(a) - 4 \left[ F^{\text{AM}}(a) \right]^2, \quad (153)$$

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

For the TiTv diff metric (Eq. 112), the support is  $\{0, 1/4, 1/2, 3/4, 1\}$ . We have already derived the probability that the TiTv diff assumes each of the values of its support (Eqs. 133-137). Therefore, we have the following distribution of the TiTv diff metric

$$f_{\text{TiTv}}(d; f_a, \gamma_0, \gamma_1, \gamma_2, \eta) = \begin{cases} (1 - f_a)^4 + 4f_a^2 (1 - f_a)^2 + f_a^4 & d = 0, \\ 4(\gamma_0 + \gamma_2) \left[ (1 - f_a)^3 f_a + f_a^3 (1 - f_a) \right] & d = 1/4, \\ 4\gamma_1 \left[ (1 - f_a)^3 f_a + f_a^3 (1 - f_a) \right] & d = 1/2, \\ 2(\gamma_0 + \gamma_2) (1 - f_a)^2 f_a^2 & d = 3/4, \\ 2\gamma_1 (1 - f_a)^2 f_a^2 & d = 1, \end{cases} \quad (154)$$

where  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu at locus  $a$ ,  $\gamma_1$  is the probability of PuPy at locus  $a$ ,  $\gamma_2$  is the probability of PyPy at locus  $a$ , and  $\eta$  is the Ti/Tv ratio (Eq. 127).

The mean and variance of this TiTv diff distribution is derived using the corresponding density function (Eq. 154). The average TiTv diff is given by

$$\text{E} \left[ d_{ij}^{\text{TiTv}}(a) \right] = (\gamma_0 + \gamma_2 + 2\gamma_1) F^{\text{TiTv}}(a) + \left[ \frac{3}{2} (\gamma_0 + \gamma_2) + 2\gamma_1 \right] G^{\text{TiTv}}(a), \quad (155)$$

where  $F^{\text{TiTv}}(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a)$ ,  $G^{\text{TiTv}}(a) = f_a^2 (1 - f_a)^2$ ,  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu at locus  $a$ ,  $\gamma_1$  is the probability of PuPy at locus  $a$ , and  $\gamma_2$  is the probability of PyPy at locus  $a$ .

The variance of the TiTv diff metric is given by

$$\begin{aligned} \text{Var} \left[ d_{ij}^{\text{TiTv}}(a) \right] &= \left[ \frac{1}{4}(\gamma_0 + \gamma_2) + \gamma_1 \right] F^{\text{TiTv}}(a) + \left[ \frac{9}{8}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] G^{\text{TiTv}}(a) \\ &\quad - \left( (\gamma_0 + \gamma_2 + 2\gamma_1) F^{\text{TiTv}}(a) + \left[ \frac{3}{2}(\gamma_0 + \gamma_2) + 2\gamma_1 \right] G^{\text{TiTv}}(a) \right)^2, \end{aligned} \quad (156)$$

where  $F^{\text{TiTv}}(a) = (1 - f_a)^3 f_a + f_a^3 (1 - f_a)$ ,  $G^{\text{TiTv}}(a) = f_a^2 (1 - f_a)^2$ ,  $f_a$  is the probability of a minor allele occurring at locus  $a$ ,  $\gamma_0$  is the probability of PuPu at locus  $a$ ,  $\gamma_1$  is the probability of PuPy at locus  $a$ , and  $\gamma_2$  is the probability of PyPy at locus  $a$ .

These novel distribution results for the projection of pairwise GWAS distances onto a single genetic variant, as well as results for the full space of  $p$  variants, can inform NPDR and other nearest-neighbor distance-based feature selection algorithms. Next we introduce our new diff metric and distribution results for time series derived correlation-based data, with a particular application to resting-state fMRI.

## 4 Time series correlation-based distance distribution

For time series correlation-based data, we consider the case where there are  $m$  correlation matrices  $A^{(p \times p)}$  (one matrix for each subject). In particular, we have in mind the application of resting-state fMRI (rs-fMRI) data. The derivations that follow, however, are relevant to all correlation-based data with the assumptions we note. The attributes in rs-fMRI are commonly Regions of Interest (ROIs), which are collections of spatially proximal voxels [22]. Correlation in their time-series activity arise between different ROIs at the voxel level or for a given brain atlas [23]. Because the attributes of interest ( $a$ ) are the ROIs themselves, we propose the following attribute projection (diff)

$$d_{ij}^{\text{ROI}}(a) = \sum_{k \neq a} |A_{ka}^{(i)} - A_{ka}^{(j)}|, \quad (157)$$

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

$$D_{ij}^{\text{fMRI}} = \sum_{a \in \mathcal{A}} d_{ij}^{\text{ROI}}(a), \quad (158)$$

which is based on Manhattan ( $q = 1$ ). This metric may be expanded to general  $q$ , but we only consider  $q = 1$ .

In order for comparisons between different correlations to be possible, we first perform a Fisher r-to-z transform on the correlations. This transformation makes the data approximately normally distributed with stabilized variance across different samples. After this transformation, we then load all of the transformed correlations into a  $p(p-1) \times m$  matrix  $X$  (Fig. 9). Each column of  $X$  represents a single instance (or subject) in rs-fMRI data. Contrary to a typical  $p \times m$  data set, each row does not represent a single attribute. Rather, each attribute (or ROI) is represented by  $p-1$  consecutive rows. The first  $p-1$  rows represent ROI<sub>1</sub>, the next  $p-1$  rows represent ROI<sub>2</sub>, and so on until the last  $p-1$  rows that represent ROI <sub>$p$</sub> . For a given column of  $X$ , we exclude pairwise correlations between an ROI and itself. Therefore, the matrix does not contain  $\hat{A}_{aa}^{(i)}$  for any  $i \in \mathcal{I}$  or  $a \in \mathcal{A}$ . Furthermore, symmetry of correlation matrices means that each column contains exactly two of each element of the upper triangle of an instance's transformed correlation matrix. For example,  $\hat{A}_{ka}^{(i)} = \hat{A}_{ak}^{(i)}$  for  $k \neq a$  and both

will be contained in a given column of  $X$  for each  $a \in \mathcal{A}$ . Based on our rs-fMRI diff (Eq. 157), the organization of  $X$  makes computation of each value of the diff very simple. In order to compute each value of the rs-fMRI diff, we just need to know the starting and ending row indices for a given ROI. Starting indices are given by

$$\text{start}_k = (k - 1)(p - 1) + 1, \quad \text{for } k = 1, 2, \dots, p$$

and ending indices are given by

$$\text{end}_k = k(p - 1), \quad \text{for } k = 1, 2, \dots, p.$$

These indices allow us to extract just the rows necessary to compute the rs-fMRI diff for a fixed ROI.

**Fig 9. Organization based on brain regions of interest (ROIs) of resting-state fMRI correlation dataset consisting of transformed correlation matrices for  $m$  subjects.** Each column corresponds to an instance (or subject)  $I_j$  and each subset of rows corresponds to the correlations for an ROI attribute ( $p$  sets). The notation  $\hat{A}_{ak}^{(j)}$  represents the r-to-z transformed correlation between attributes (ROIs)  $a$  and  $k \neq a$  for instance  $j$ .

We further transform the data matrix  $X$  by standardizing so that each of the  $m$  columns has zero mean and unit variance. Therefore, the data in matrix  $X$  are approximately standard normal. Recall that the mean (Eq. 41) and variance (Eq. 42) of the Manhattan ( $L_1$ ) distance distribution for standard normal data are  $\frac{2p}{\sqrt{\pi}}$  and  $\frac{2(\pi-2)p}{\pi}$ , respectively. This allows us to easily derive the expected pairwise distance between instances  $i, j \in \mathcal{I}$  in rs-fMRI data as follows

$$\begin{aligned} \mathbb{E}(D_{ij}^{\text{fMRI}}) &= \mathbb{E} \left( \sum_{a \in \mathcal{A}} d_{ij}^{\text{ROI}}(a) \right) \\ &= \mathbb{E} \left( \sum_{a \in \mathcal{A}} \sum_{k \neq a} \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\ &= \sum_{a \in \mathcal{A}} \sum_{k \neq a} \mathbb{E} \left( \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\ &= \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{2}{\sqrt{\pi}} \\ &= \frac{2p(p-1)}{\sqrt{\pi}}. \end{aligned} \tag{159}$$

The expected pairwise rs-fMRI distance (Eq. 159) grows on the order of  $p(p-1)$ , which is the total number of transformed pairwise correlations in each column of  $X$  (Fig. 9). This is similar to the case of a typical  $m \times p$  data matrix in which the data is standard normal and Manhattan distances are computed between instances.

We first derive the variance of the rs-fMRI distance by making an independence assumption with respect to the magnitude differences  $|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|$  for all  $k \neq a \in \mathcal{A}$ . We observe empirically that this assumption gives a reasonable estimate of the actual variance of rs-fMRI distances in simulated data, but there is a consistent discrepancy between predicted and simulated variances. We begin our derivation of the variance of rs-fMRI distances by assuming that cross-covariances between the diffs of different pairs of ROIs are negligible. This allows us to determine the relationship between the predicted variance under the independence assumption and the simulated variance. We proceed by applying the variance operator linearly as follows

$$\begin{aligned}
\text{Var}(D_{ij}^{\text{fMRI}}) &= \text{Var}\left(\sum_{a \in \mathcal{A}} d_{ij}^{\text{ROI}}(a)\right) \\
&= \text{Var}\left(\sum_{a \in \mathcal{A}} \sum_{k \neq a} |\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|\right) \\
&= \sum_{a \in \mathcal{A}} \sum_{k \neq a} \text{Var}\left(|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|\right) \\
&= \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{2(\pi - 2)}{\pi} \\
&= \frac{2(\pi - 2)(p - 1)p}{\pi}.
\end{aligned} \tag{160}$$

Similar to the case of an  $m \times p$  data matrix containing standard normal data, we have an rs-fMRI distance variance that grows on the order of  $p(p-1)$ , which is the total number of pairwise associations in a column of data matrix  $X$  (Fig. 9). Therefore, the expected rs-fMRI distance (Eq. 159) and the variance of the rs-fMRI distance (Eq. 160) increase on the same order.

The independence assumption used to derive the variance of our rs-fMRI distance metric (Eq. 160) is not satisfied because a single value of the diff (Eq. 157) includes the same fixed ROI,  $a$ , for each term in the sum for all  $k \neq a$ . Therefore, the linear application of the variance operator we have previously employed does not account for the additional cross-covariance that exists. However, we have seen empirically that the theoretical variance of the distance we computed for the rs-fMRI distance metric (Eq. 160) still reasonably approximates the sample variance, there is a slight discrepancy between our theoretical rs-fMRI distance metric variance (Eq. 160) and the sample variance. More precisely, the formula we have given for the variance (Eq. 160) consistently underestimates the sample variance of the rs-fMRI distance. To adjust for this discrepancy, we determine a corrected formula by assuming that there is dependence between the terms of the rs-fMRI diff and estimate the cross-covariance between rs-fMRI diffs of different pairs of ROIs.

We begin the derivation of our corrected formula by writing the variance as a two-part sum, where the first term in the sum involves the variance of the magnitude difference  $|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|$  and then second term involves the cross-covariance of the rs-fMRI diff for distinct pairwise ROI-ROI associations. This formulation is implied in our previous derivation of the variance, but our independence assumption allowed us to assume that all terms in the second part of the two-part sum were zero. Our

formulation of the variance is given by the following

1074

$$\begin{aligned}
\text{Var}(D_{ij}^{\text{fMRI}}) &= \text{Var} \left( \sum_{a \in \mathcal{A}} \sum_{k \neq a} \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\
&= \sum_{a=1}^{p-1} \text{Var} \left( \sum_{k=a+1}^p 2 \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left( \sum_{k=a+1}^p 2 \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right|, \sum_{s=r+1}^p 2 \left| \hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)} \right| \right) \\
&= \sum_{a=1}^{p-1} \sum_{k=a+1}^p \text{Var} \left( 2 \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left( \sum_{k=a+1}^p 2 \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right|, \sum_{s=r+1}^p 2 \left| \hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)} \right| \right) \quad (161) \\
&= \sum_{a=1}^{p-1} \sum_{k=a+1}^p \frac{4(\pi-2)}{\pi} \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left( \sum_{k=a+1}^p 2 \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right|, \sum_{s=r+1}^p 2 \left| \hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)} \right| \right) \\
&= \frac{2p(\pi-2)(p-1)}{\pi} \\
&\quad + 2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left( \sum_{k=a+1}^p 2 \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right|, \sum_{s=r+1}^p 2 \left| \hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)} \right| \right).
\end{aligned}$$

In order to have a formula in terms of the number of ROIs  $p$  only, we estimate the double sum on the right-hand side of the equation of rs-fMRI distance variance (Eq. 161). Through simulation, it can be seen that the difference between 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

1075

1076

1077

1078

1079

$$S_{D_{ij}}^2 - \frac{2p(\pi-2)(p-1)}{\pi} = \beta_1 p^2 + \beta_0 p. \quad (162)$$

where  $\beta_0$  and  $\beta_1$  are the coefficients we must estimate in order to approximate the cross-covariance term in the right-hand side of the rs-fMRI distance variance equation (Eq. 161).

1080

1081

1082

The coefficient estimates found through least squares fitting are  $\beta_1 = -\beta_0 \approx 0.08$ . These estimates allow us to arrive at a functional form for the double sum in the right-hand side of the rs-fMRI distance variance equation (Eq. 161) that is proportional to  $\frac{2p(\pi-2)(p-1)}{\pi}$ . That is, we have the following formula for approximating the double sum

1083

1084

1085

1086

1087

$$2 \sum_{a=1}^{p-1} \sum_{r=a+1}^{p-1} \text{Cov} \left( \sum_{k=a+1}^p 2 \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right|, \sum_{s=r+1}^p 2 \left| \hat{A}_{rs}^{(i)} - \hat{A}_{rs}^{(j)} \right| \right) \approx \frac{p(\pi-2)(p-1)}{4\pi}. \quad (163)$$

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

1088

1089

$$\text{Var}(D_{ij}^{\text{fMRI}}) \approx \frac{9p(\pi-2)(p-1)}{4\pi}. \quad (164)$$

With the mean (Eq. 159) and variance (Eq. 164) estimates, we have the following asymptotic distribution for rs-fMRI distances

$$D_{ij}^{\text{fMRI}} \sim \mathcal{N}\left(\frac{2p(p-1)}{\sqrt{\pi}}, \frac{9p(\pi-2)(p-1)}{4\pi}\right). \quad (165)$$

#### 4.1 Max-min normalized time series correlation-based distance distribution

Previously (Section 2.4) we determined the asymptotic distribution of the sample maximum of size  $m$  from a standard normal distribution. We can naturally extend these results to our transformed rs-fMRI data because  $X$  (Fig. 9) is approximately standard normal. We proceed with the definition of the max-min normalized rs-fMRI pairwise distance.

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

$$D_{ij}^{\text{fMRI}*} = \sum_{a \in \mathcal{A}} \sum_{k \neq a} \frac{|A_{ak}^{(i)} - A_{ak}^{(j)}|}{\max(a) - \min(a)}. \quad (166)$$

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

$$\mathbb{E}(X_a^{\max} - X_a^{\min}) = 2\mu_{\max}^{(1)}(m, p) = 2 \left[ \frac{\log(\log(2))}{\Phi^{-1}\left(\frac{1}{m(p-1)}\right)} - \Phi^{-1}\left(\frac{1}{m(p-1)}\right) \right]. \quad (167)$$

The variance can be esimated with the following

$$\text{Var}(X_a^{\max} - X_a^{\min}) = \frac{\pi^2}{6\log[m(p-1)]}. \quad (168)$$

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. 159 and 164. Using the formulas for the mean and variance of the max-min normalized distance distribution given in Eq. 92, we have the following asymptotic distribution for the max-min normalized rs-fMRI distances

$$D_{ij}^{\text{fMRI}*} \sim \mathcal{N}\left(\frac{\mu_{D_{ij}^{\text{fMRI}}}}{2\mu_{\max}^{(1)}(m, p)}, \frac{6\sigma_{D_{ij}^{\text{fMRI}}}^2 \log[m(p-1)]}{\pi^2 + 24 \left[\mu_{\max}^{(1)}(m, p)\right]^2 \log[m(p-1)]}\right). \quad (169)$$

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

Just as in previous sections (Sections. 2.2.3 and 3.4), we now derive the distribution of our rs-fMRI diff metric (Eq. 157). Unlike what we have seen in previous sections, we do not derive the exact distribution for this diff metric. We have determined empirically that the rs-fMRI diff is approximately normal. Although the rs-fMRI diff is a sum of  $p-1$  magnitude differences, the Classical Central Limit Theorem does not apply because of the dependencies that exist between the terms of the sum. Examination of histograms and quantile-quantile plots of simulated values of the rs-fMRI diff easily indicate that the normality assumption is safe. Therefore, we derive the mean and



variance of the approximately normal distribution of the rs-fMRI diff. As we have seen previously, this normality assumption is reasonable even for small values of  $p$ .

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

$$\begin{aligned} \mathbb{E} \left[ d_{ij}^{\text{ROI}}(a) \right] &= \mathbb{E} \left( \sum_{k \neq a} \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\ &= \sum_{k \neq a} \mathbb{E} \left( \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\ &= \sum_{k \neq a} \frac{2}{\sqrt{\pi}} \\ &= \frac{2(p-1)}{\sqrt{\pi}}, \end{aligned} \tag{170}$$

where  $a$  is a single fixed ROI.

Considering the variance of the rs-fMRI diff metric, we have two estimates. The first estimate uses the variance operator in a linear fashion, while the second will simply be a direct implication of the corrected formula of the variance of rs-fMRI pairwise distances (Eq. 164). Our first estimate is derived as follows

$$\begin{aligned} \text{Var} \left[ d_{ij}^{\text{ROI}}(a) \right] &= \text{Var} \left( \sum_{k \neq a} \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\ &= \sum_{k \neq a} \text{Var} \left( \left| \hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)} \right| \right) \\ &= \sum_{k \neq a} \frac{2(\pi-2)}{\pi} \\ &= \frac{2(\pi-2)(p-1)}{\pi}, \end{aligned} \tag{171}$$

where  $a$  is a single fixed ROI.

Using the corrected rs-fMRI distance variance formula (Eq. 164), our second estimate of the rs-fMRI diff variance is given directly by the following

$$\text{Var} \left[ d_{ij}^{\text{ROI}}(a) \right] = \frac{9(\pi-2)(p-1)}{4\pi}, \tag{172}$$

where  $a$  is a single fixed ROI.

Empirically, the first estimate (Eq. 171) of the variance of our rs-fMRI diff is closer to the sample variance than the second estimate (Eq. 172). This is due to fact that we are considering only a fixed ROI  $a \in \mathcal{A}$ , so the cross-covariance between the magnitude differences  $|\hat{A}_{ak}^{(i)} - \hat{A}_{ak}^{(j)}|$  for different pairs of ROIs ( $a$  and  $k \neq a$ ) is negligible here. When considering all ROIs  $a \in \mathcal{A}$ , these cross-covariances are no longer negligible. Using the first variance estimate (Eq. 171) and the estimate of the mean (Eq. 170), we have the following asymptotic distribution of the rs-fMRI diff

$$d_{ij}^{\text{ROI}}(a) \sim \mathcal{N} \left( \frac{2(p-1)}{\sqrt{\pi}}, \frac{2(\pi-2)(p-1)}{\pi} \right), \tag{173}$$

where  $a$  is a single fixed ROI.

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

Substituting the non-normalized mean (Eq. 159) into the equation for the mean of the max-min normalized rs-fMRI metric (Eq. 169), we have the following

$$\begin{aligned} E(D_{ij}^{\text{fMRI}*}) &= \frac{\mu_{D_{ij}^{\text{fMRI}}}^{(1)}}{2\mu_{\max}^{(1)}(m, p)} \\ &= \frac{p(p-1)}{\sqrt{\pi}\mu_{\max}^{(1)}(m, p)}, \end{aligned} \quad (174)$$

where  $\mu_{\max}^{(1)}(m, p)$  (Eq. 167) is the expected maximum of a single ROI in a data set with  $m$  instances and  $p$  ROIs.

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

$$\begin{aligned} \text{Var}(D_{ij}^{\text{fMRI}*}) &= \frac{6\sigma_{D_{ij}^{\text{fMRI}}}^2 \log[m(p-1)]}{\pi^2 + 24 \left[ \mu_{\max}^{(1)}(m, p) \right]^2 \log[m(p-1)]} \\ &= \frac{27(\pi-2)\log[m(p-1)](p-1)p}{2\pi \left( \pi^2 + 24 \left[ \mu_{\max}^{(1)}(m, p) \right]^2 \log[m(p-1)] \right)}, \end{aligned} \quad (175)$$

where  $\mu_{\max}^{(1)}(m, p)$  (Eq. 167) is the expected maximum of a single ROI in a data set with  $m$  instances and  $p$  ROIs.

We summarize the moment estimates for the rs-fMRI metrics for correlation-based data derived from time series (Fig. 10). We organize this summary by standard and attribute range-normalized rs-fMRI distance metric, statistic (mean or variance), and asymptotic formula.

**Fig 10. Asymptotic means and variances for the new standard (Eq. 158) and max-min normalized (Eq. 166) rs-fMRI distance metrics.**

## 5 Comparison of theoretical and sample moments

We compare our analytical asymptotic estimates of sample moments for distributions of pairwise distances in high attribute dimension by generating random data for various dimensions  $m$  and  $p$  (Fig. 11). We fix  $m = 100$  samples and compute Manhattan (Eq. 1) distance matrices from standard normal data for  $p = 1000, 2000, 3000, 4000$ , and  $5000$  attributes. For each value of  $p$ , we generate 20 random datasets and compute the mean and standard deviation of pairwise distances. We then average these 20 simulated means and standard deviations. For comparison, we compute the theoretical moments (Eqs. 41 and 42) for each value of  $p$  and fixed  $m = 100$  from the theoretical formulas. Scatter plots of theoretical versus simulated mean (Fig. 11A) and theoretical versus simulated standard deviation (Fig. 11B) indicate that our theoretical asymptotic formulas for sample moments are reliable for both large and relatively small numbers of attributes. For other combinations of data type, distance metric, sample size  $m$ , and number of attributes  $p$ , we find similar agreement between theoretical formulas and simulated moments (not shown).

**Fig 11. Comparison of theoretical and sample moments of Manhattan (Eq. 1) distances in standard normal data. (A) Scatter plot of theoretical vs simulated mean Manhattan distance (Eq. 41).** Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard normal data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. **(B) Scatter plot of theoretical vs simulated standard deviation of Manhattan (Eq. 1) distance (Eq. 42).** These standard deviations come from the same random distance matrices for which mean distance was computed for **A**. Both theoretical mean and standard deviation approximate the simulated moments quite well.

## 6 Effects of correlation on distances

All of the derivations presented in previous sections were for the cases where there is no correlation between instances or between attributes. We assumed that any pair  $(X_{ia}, X_{ja})$  of data points for instances  $i$  and  $j$  and fixed attribute  $a$  were independent and identically distributed. This was assumed in order to determine asymptotic estimates in null data. That is, data with no main effects, interaction effects, or pairwise correlations between attributes. Within this simplified context, our asymptotic formulas for distributional moments are reliable. However, in real data are numerous statistical effects that impact distance distributional properties. We find that deviation from normality is caused primarily by large magnitude pairwise correlation between attributes. Pairwise attribute correlation can be the result of main effects, where attributes have different within-group means. On the other hand, there could be an underlying interaction network in which there are strong associations between attributes. If attributes are differentially correlated between phenotype groups, then interactions exist that change 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 attribute correlation.

### 6.1 Continuous data

Without loss of generality, suppose we have  $X^{(m \times p)}$  where  $X_{ia} \sim \mathcal{N}(0, 1)$  for all  $i = 1, 2, \dots, m$  and  $a = 1, 2, \dots, p$ , and let  $m = p = 100$ . We consider only the  $L_2$  (Euclidean) metric (Eq. 1,  $q = 2$ ). We explore the effects of correlation on these distances by generating simulated data sets with increasing strength of pairwise attribute correlation and then plotting the density curve of the induced distances (Fig. 12A). Deviation from normality in the distance distribution is directly related to the average absolute pairwise correlation that exists in the simulated data. This measure is given by

$$\bar{r}_{\text{abs}} = \frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j>i} r_{ij} \quad (176)$$

where  $r_{ak}$  is the correlation between attributes  $a, k \in \mathcal{A}$  across all instances  $m$ . Distances generated on data without correlation closely approximate a Gaussian. The mean (Eq. 53) and variance (Eq. 52) of the uncorrelated distance distribution are given by substituting  $p = 100$  for the mean. As  $\bar{r}_{\text{abs}}$  increases, positive skewness and increased variability in distances emerges. The predicted and sample means, however, are approximately the same between correlated and uncorrelated distances due to linearity of the expectation operator. Because of the dependencies between attributes, the predicted variance of 1 for  $L_2$  on standard normal data obviously no longer holds.

In order to introduce a controlled level of correlation between attributes, we created correlation matrices based on a random graph with specified connection probability, where attributes correspond to the vertices in each graph. We assigned high correlations to connected attributes 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  $X^{\text{corr}}$

$$X^{\text{corr}} = XU^T. \quad (177)$$

The new data matrix  $X^{\text{corr}}$  has approximately the same correlation structure as the randomly generated correlation matrix created from a random graph.

## 6.2 GWAS data

Analogous to the previous section, we explore the effects of pairwise attribute correlation in the context of GWAS data. Without loss of generality, we let  $m = p = 100$  and consider only the TiTv metric (Eq. 115). To create correlated GWAS data, we first generated standard normal data with random correlation structure, just as in the previous section. We then applied the standard normal cumulative distribution function (CDF) to this correlated data in order transform the correlated standard normal variates into uniform data with preserved correlation structure. We then subsequently applied the inverse binomial CDF to the correlated uniform data with random success probabilities  $f_a$  for all  $a \in \mathcal{A}$ . Each attribute  $a \in \mathcal{A}$  corresponds to an individual SNP in the data matrix. The resulting GWAS data set is binomial with  $n = 2$  trials and has roughly the same correlation matrix as the original correlated standard normal data with which we started. Average absolute pairwise correlation  $\bar{r}_{\text{abs}}$  induces positive skewness in GWAS data at lower levels than in correlated standard normal data (Fig. 12B). This could have important implications in nearest neighborhoods in NPDR and similar methods.

## 6.3 Time-series derived correlation-based datasets

For our correlation data-based metric (Eq. 158), we consider additional effects of correlation between features. Without loss of generality, we let  $m = 100$  and  $p = 30$ . We show an illustration of the effects of correlated features in this context (Fig. 12C). Based on the density estimates, it appears that correlation between features introduces positive skewness at low values of  $\bar{r}_{\text{abs}}$ . We introduced correlation to the transformed data matrix (Fig. 9) with the cholesky method used previously.

**Fig 12. Distance densities from uncorrelated vs correlated bioinformatics data.** (A) Euclidean distance densities for random normal data with and without correlation. Correlated data was created by multiplying random normal data by upper-triangular Cholesky factor from randomly generated correlation matrix. We created correlated data for average absolute pairwise correlation (Eq. 176)  $\bar{r}_{\text{abs}} = 0.105, 0.263, 0.458$ , and  $0.612$ . (B) TiTv distance densities for random binomial data with and without correlation. Correlated data was created by first generating correlated standard normal data using the Cholesky method from (A). Then we applied the standard normal CDF to create correlated uniformly distributed data, which was then transformed by the inverse binomial CDF with  $n = 2$  trials and success probabilities  $f_a$  for all  $a \in \mathcal{A}$ . (C) Time series correlation-based distance densities for random rs-fMRI data (Fig. 9) with and without additional pairwise feature correlation. Correlation was added to the transformed rs-fMRI data matrix (Fig. 9) using the Cholesky algorithm from (A).

## 7 Discussion

Nearest-neighbor distance-based feature selection is a class of methods that are relatively simple to implement, and they perform well at detecting interaction effects in high dimensional data. Theoretical analysis of the limiting behavior of distance distributions for various data types and dimensions may lead to improved hyperparameter estimates of these feature selection methods. Furthermore, these theoretical results may help guide the choice of distance metric for a given dataset. Most often, distance-based feature selection methods use the  $L_q$  metric (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  $L_2$  for standard normal data is on the order of  $p$  (Eq. 41 and Fig. 3) and  $\sqrt{p}$  (Eq. 51 and Fig. 3), respectively. In addition,  $L_1$  and  $L_2$  on standard normal data have asymptotic variances on the order of  $p$  and 1, respectively (Eqs. 42 and 52).

These results can inform the choice of  $L_1$  or  $L_2$  depending context. 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  $L_2$ , the asymptotic distribution ( $\mathcal{N}(\sqrt{2p} - \bar{1}, 1)$ ) indicates that the limiting  $L_2$  distribution can be thought of simply as a positive translation of the standard normal distribution ( $\mathcal{N}(0, 1)$ ). The  $L_2$  distribution ( $\mathcal{N}(\sqrt{2p} - \bar{1}, 1)$ ) also indicates that most neighbors are contained in a thin shell far from the instance in high dimension ( $p \gg 1$ ). 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 attribute dimension  $p$ . This  $L_1$  dispersion could 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 attribute dimension  $p$ , which in turn could affect the quality of selected attributes. 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 our analysis, we 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 (Eq. 115) and time series correlation-based data (Eqs. 158 and 166) like resting-state fMRI. We also extended the asymptotic results for the mean and variance of the attribute range-normalized  $L_q$  (max-min) distance for standard normal (Eq. 92) and standard uniform (Eq. 100) data using extreme value theory. Our derivations provide an important reference for those using nearest-neighbor feature selection or classification methods in common bioinformatics data.

In this work, we 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 application to resting-state fMRI data. In order for this to be possible, we created a new metric (Eq. 157) that allows us to use regions of interest (ROIs) as attributes. Not all ROIs will be relevant to a particular phenotype in case-control studies, and nearest-neighbor feature selection would be a useful tool to determine important ROIs due to interactions and to help elucidate the network structure of the brain as it relates to the phenotype of interest.

The recently introduced transition-transversion metric (Eq. 112) 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 (Eq. 110) and AM (Eq. 111) metrics, provides an important reference to aid in neighborhood parameter selection in this context. We have also shown how the Ti/TV ratio  $\eta$  (Eq. 127) and minor allele frequency (or success

probability)  $f_a$  affects these discrete distances. For the GM and AM metrics, the distance is solely determined by the minor allele frequencies because the genotype encoding is not taken into account. We showed how both minor allele frequency and Ti/Tv ratio uniquely affects the TiTv distance (Figs. 7A and 7C). 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.

In addition to asymptotic  $L_q$  distance distributions, we have also provided the exact distributions for the one-dimensional projection of the  $L_q$  distance onto individual attributes (Sections. 2.2.3, 3.4, and 4.2). These distributions are important for all nearest-neighbor distance-based feature selection algorithms, such as Relief or NPDR, because the  $L_q$  distance is a function of the one-dimensional attribute projection (diff). In particular, these projected distance distributions are important for improving inference for predictors in NPDR, which are one-dimensional attribute projections.

Correlations between attributes and instances can cause significant deviations from the asymptotic results for uncorrelated data we have derived in this work. To illustrate this behavior, we showed how strong correlations lead to positive skewness in the distance distribution of random normal, binomial, and rs-fMRI data (Figs. 12A, 12B, and 12C). Pairwise correlation between attributes does not change the average distance, so our asymptotic results for uncorrelated data also apply when attributes are not independent. In contrast, the sample variance of distances deviates from the uncorrelated case substantially as the average absolute pairwise attribute correlation increases (Eq. 176). For fixed-radius neighborhood methods, this deviation increases the probability of including neighbors for a given instance and may reduce the power to detect interactions. The increased variability for distances with correlated data may inform the choice of metric and optimization of neighborhoods in nearest-neighbor feature selection.

## 8 Supporting information

**S1 Fig. 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.

**S2 Fig. 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.

**S3 Fig. Convergence to Gaussian for max-min normalized 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.

**S4 Fig. Convergence to Gaussian for GM distances for simulated binomial GWAS data with  $m = 100$  instances and  $p = 10, 100, 1000$ , and 10000 attributes.** The average MAF was set to 0.205 for all simulations. Convergence to Gaussian occurs more gradually with increasing  $p$  than in continuous data. Significant convergence seems to occur when  $p \geq 1000$ , however, this is actually a relatively small number of features in the context of GWAS. Considering a realistic number of features for GWAS, the normality assumption of GM distances holds. This metric has the slowest convergence to Gaussian among all we have considered. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

**S5 Fig. Convergence to Gaussian for AM distances for simulated binomial GWAS data with  $m = 100$  instances and  $p = 10, 100, 1000$ , and 10000 attributes.** The average MAF was set to 0.205 for all simulations. Convergence to Gaussian occurs more gradually with increasing  $p$  than in continuous data. Significant convergence seems to occur when  $p \geq 1000$ , however, this is actually a relatively small number of features in the context of GWAS. Considering a realistic number of features for GWAS, the normality assumption of AM distances holds. This metric has the slightly faster convergence to Gaussian than the GM metric, which is probably due to the fact that the AM metric has one more value in its range (e.g.,  $1/2$ ). P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

**S6 Fig. Convergence to Gaussian for TiTv distances for simulated binomial GWAS data with  $m = 100$  instances and  $p = 10, 100, 1000$ , and 10000 attributes.** The average MAF was set to 0.205 for all simulations and the Ti/Tv ratio ( $\eta$ ) was set to 2. Convergence to Gaussian occurs more gradually with increasing  $p$  than in continuous data. Significant convergence seems to occur when  $p \geq 1000$ , however, this is actually a relatively small number of features in the context of GWAS. Considering a realistic number of features for GWAS, the normality assumption of TiTv distances holds. This metric has the significantly faster convergence to Gaussian than the AM metric, which is probably due to the fact that the TiTv metric contains 2 more values in its range (e.g.,  $1/4$  &  $3/4$ ). P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

**S7 Fig. Convergence to Gaussian for rs-fMRI distances for simulated correlation matrices with  $m = 100$  instances and  $p = 10, 50, 150$ , and 300 attributes (or ROIs).** Correlation matrices were generated for each instance from random normal  $m \times p$  data matrices. Each correlation matrix was then stretched out into a long vector, Fisher r-to-z transformed, stored in a  $p(p-1) \times m$  matrix, and standardized so that the  $m$  columns are mean 0 and unit variance. Convergence to Gaussian occurs very rapidly for this data because the dimensions are larger than a typical  $m \times p$  data set. The large attribute dimension  $p(p-1)$  means that there are significantly more terms in each sum to compute pairwise distances. Therefore, Classical Central Limit Theorem dictates that distances in this context will be closer to Gaussian. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

**S8 Fig. Convergence to Gaussian for max-min normalized rs-fMRI distances for simulated correlation matrices with  $m = 100$  instances and  $p = 10, 50, 150$ , and 300 attributes (or ROIs).** Correlation matrices were generated for each instance from random normal  $m \times p$  data matrices. Each correlation matrix

was then stretched out into a long vector, Fisher r-to-z transformed, stored in a  $p(p-1) \times m$  matrix, and standardized so that the  $m$  columns are mean 0 and unit variance. Convergence to Gaussian occurs approximately as rapidly as the standard rs-fMRI metric. Just as in the standard rs-fMRI metric, the large attribute dimension  $p(p-1)$  means that there are significantly more terms in each sum to compute pairwise distances. Therefore, Classical Central Limit Theorem dictates that distances in this context will be closer to Gaussian. P values from Shapiro-Wilk test, where the null hypothesis is a Gaussian distribution.

**S9 Fig. Comparison of theoretical and simulated moments of Manhattan distances in standard normal data.** (A) Scatter plot of theoretical vs simulated mean Manhattan distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard normal data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of Manhattan distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S10 Fig. Comparison of theoretical and simulated moments of max-min normalized Manhattan distances in standard normal data.** (A) Scatter plot of theoretical vs simulated mean max-min normalized Manhattan distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard normal data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of max-min normalized Manhattan distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S11 Fig. Comparison of theoretical and simulated moments of Euclidean distances in standard normal data.** (A) Scatter plot of theoretical vs simulated mean Euclidean distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard normal data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of Euclidean distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Theoretical and simulated means lie approximately on the identity line because the mean is proportional to attribute dimension  $p$ . Theoretical standard deviation is constant, which is why each horizontal coordinate is the same for  $p = 1000, 2000, 3000, 4000$ , and 5000. The variation in sample standard deviation of Euclidean distance is quite small, so each simulated moment is clustered about 1.

**S12 Fig. Comparison of theoretical and simulated moments of max-min normalized Euclidean distances in standard normal data.** (A) Scatter plot of theoretical vs simulated mean max-min normalized Euclidean distance. Each point



represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard normal data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. **(B)** Scatter plot of theoretical vs simulated standard deviation of max-min normalized Euclidean distance. These standard deviations come from the same random distance matrices for which mean distance was computed for **A**. Theoretical and simulated means lie approximately on the identity line because the mean is proportional to  $\sqrt{p}$ . Theoretical standard deviation is a function of the fixed attribute dimension  $m$ , which is why each horizontal coordinate is the same for  $p = 1000, 2000, 3000, 4000$ , and  $5000$ . The variation in sample standard deviation of max-min normalized Euclidean distance is quite small, so each simulated moment is clustered about the theoretical value that depends on  $m$ .

**S13 Fig. Comparison of theoretical and simulated moments of Manhattan distances in standard uniform data.** **(A)** Scatter plot of theoretical vs simulated mean Manhattan distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard uniform data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. **(B)** Scatter plot of theoretical vs simulated standard deviation of Manhattan distance. These standard deviations come from the same random distance matrices for which mean distance was computed for **A**. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S14 Fig. Comparison of theoretical and simulated moments of Euclidean distances in standard uniform data.** **(A)** Scatter plot of theoretical vs simulated mean Euclidean distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard uniform data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. **(B)** Scatter plot of theoretical vs simulated standard deviation of Euclidean distance. These standard deviations come from the same random distance matrices for which mean distance was computed for **A**. Theoretical and simulated means lie approximately on the identity line because the mean is proportional to attribute dimension  $p$ . Theoretical standard deviation is constant, which is why each horizontal coordinate is the same for  $p = 1000, 2000, 3000, 4000$ , and  $5000$ . The variation in sample standard deviation of Euclidean distance is quite small, so each simulated moment is clustered about  $7/120$ .

**S15 Fig. Comparison of theoretical and simulated moments of max-min normalized Euclidean distances in standard uniform data.** **(A)** Scatter plot of theoretical vs simulated mean max-min normalized Euclidean distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from standard uniform data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. **(B)** Scatter plot of theoretical vs simulated standard deviation of max-min normalized Euclidean distance. These standard deviations come from the same random distance matrices for which mean distance was computed for **A**. Theoretical and simulated means lie approximately on the identity line

because the mean is proportional to  $\sqrt{p}$ . Theoretical standard deviation is a function of the fixed attribute dimension  $m$ , which is why each horizontal coordinate is the same for  $p = 1000, 2000, 3000, 4000$ , and  $5000$ . The variation in sample standard deviation of max-min normalized Euclidean distance is quite small, so each simulated moment is clustered about the theoretical value that depends on  $m$ .

**S16 Fig. Comparison of theoretical and simulated moments of GM distances in binomial GWAS data.** (A) Scatter plot of theoretical vs simulated mean GM distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from binomial GWAS data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of GM distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S17 Fig. Comparison of theoretical and simulated moments of AM distances in binomial GWAS data.** (A) Scatter plot of theoretical vs simulated mean AM distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from binomial GWAS data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of AM distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S18 Fig. Comparison of theoretical and simulated moments of TiTv distances in binomial GWAS data.** For each simulated data set, the average MAF was set to be 0.205 and the Ti/Tv ratio ( $\eta$ ) was fixed to be 2. (A) Scatter plot of theoretical vs simulated mean TiTv distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from binomial GWAS data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of TiTv distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S19 Fig. Comparison of theoretical and simulated moments of rs-fMRI distances from random correlation matrices.** For each instance, we generated a  $p \times p$  correlation matrix from a random  $m \times p$  standard normal data set. We then stretched out each correlation matrix into a long vector, Fisher r-to-z transformed the correlations, stored the vector in a column of a large  $p(p-1) \times m$  matrix, and then standardized columns to be mean 0 and unit variance. (A) Scatter plot of theoretical vs simulated mean rs-fMRI distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from rs-fMRI data and computed the average simulated pairwise distance from the 20

iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of rs-fMRI distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S20 Fig. Comparison of theoretical and simulated moments of max-min normalized rs-fMRI distances from random correlation matrices.** For each instance, we generated a  $p \times p$  correlation matrix from a random  $m \times p$  standard normal data set. We then stretched out each correlation matrix into a long vector, Fisher r-to-z transformed the correlations, stored the vector in a column of a large  $p(p-1) \times m$  matrix, and then standardized columns to be mean 0 and unit variance. (A) Scatter plot of theoretical vs simulated mean max-min normalized rs-fMRI distance. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 distance matrices from rs-fMRI data and computed the average simulated pairwise distance from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of max-min normalized rs-fMRI distance. These standard deviations come from the same random distance matrices for which mean distance was computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S21 Fig. Comparison of theoretical and simulated moments of rs-fMRI diff metric from random correlation matrices.** For each instance, we generated a  $p \times p$  correlation matrix from a random  $m \times p$  standard normal data set. We then stretched out each correlation matrix into a long vector, Fisher r-to-z transformed the correlations, stored the vector in a column of a large  $p(p-1) \times m$  matrix, and then standardized columns to be mean 0 and unit variance. (A) Scatter plot of theoretical vs simulated mean rs-fMRI diff. Each point represents a different number of attributes  $p$ . For each value of  $p$  we fixed  $m = 100$  and generated 20 diff metric values from the rs-fMRI data and computed the average simulated diff from the 20 iterations. The corresponding theoretical mean was then computed for each value of  $p$  for comparison. The dashed line represents the identity (or  $y = x$ ) line for reference. (B) Scatter plot of theoretical vs simulated standard deviation of rs-fMRI diff. These standard deviations come from the same random diff values from which mean diffs were computed for A. Both theoretical mean and standard deviation approximate the simulated moments quite well.

**S22 Fig. Density curves for one-dimensional projected distances (diffs) onto a fixed attribute  $a$  for standard normal and standard uniform data.** (A) Density curves for the distribution of attribute diff in standard normal data for  $q = 1, 2, 3, 4$ , and 5. This density is that of a Generalized Gamma distribution. For  $q = 1$ , this is also known as a half-normal distribution. (B) Density curves for the distribution of attribute diff in standard uniform data for  $q = 1, 2, 3, 4$ , and 5. This density is that of a Kumaraswamy distribution. For  $q = 1$ , this is also known as a triangular distribution.

**S23 Fig. One-dimensional projected GM distance (diff) onto an attribute vs minor allele frequency (MAF).** For each possible value of the GM diff ( $d = 0, 1$ ), the exact density of the GM diff is plotted for all possible values of MAF. The expected value of MAF at a particular locus  $a$  is  $f_a$  for all  $a \in \mathcal{A}$ , where  $f_a$  is the probability of a

minor allele occurring at locus  $a$ . For each element  $X_{ia}$  of the data matrix for a fixed attribute  $a$ , we have  $X_{ia} \sim \mathcal{B}(2, f_a)$ . Depending on the MAF, the frequency of GM diff taking on a value of 0 or 1 changes. For small MAF, the GM diff will be 0 most often. As MAF increases beyond 0.5, the minor allele switches.

**S24 Fig. One-dimensional projected AM distance (diff) onto an attribute vs minor allele frequency (MAF).** For each possible value of the AM diff ( $d = 0, 1/2, 1$ ), the exact density of the AM diff is plotted for all possible values of MAF. The expected value of MAF at a particular locus  $a$  is  $f_a$  for all  $a \in \mathcal{A}$ , where  $f_a$  is the probability of a minor allele occurring at locus  $a$ . For each element  $X_{ia}$  of the data matrix for a fixed attribute  $a$ , we have  $X_{ia} \sim \mathcal{B}(2, f_a)$ . Depending on the MAF, the frequency of AM diff taking on a value of 0,  $1/2$ , or 1 changes. For small MAF, the AM diff will be 0 most often. For large MAF ( $\approx 0.5$ ), the AM diff will be mostly  $1/2$  with 1 being the second most common value. The least common value of the AM diff is 1. As MAF increases beyond 0.5, the minor allele switches.

**S25 Fig. One-dimensional projected TiTv distance (diff) onto an attribute vs minor allele frequency (MAF).** For each possible value of the TiTv diff ( $d = 0, 1/4, 1/2, 3/4, 1$ ), the exact density of the TiTv diff is plotted for all possible values of MAF. The Ti/Tv ratio  $\eta$  was fixed to be 2. The expected value of MAF at a particular locus  $a$  is  $f_a$  for all  $a \in \mathcal{A}$ , where  $f_a$  is the probability of a minor allele occurring at locus  $a$ . For each element  $X_{ia}$  of the data matrix for a fixed attribute  $a$ , we have  $X_{ia} \sim \mathcal{B}(2, f_a)$ . Depending on the MAF, the frequency of TiTv diff taking on a value of 0,  $1/4$ ,  $1/2$ ,  $3/4$ , or 1 changes. The density of the TiTv diff for  $d = 1/4, 1/2$ , and  $3/4$  has the same resulting curve as a function of MAF. The most common value of TiTv diff at any MAF is 0, the second most common is  $1/4$ ,  $1/2$ , or  $3/4$ , and the least common is 1. As MAF increases beyond 0.5, the minor allele switches.

## References

1. Urbanowicz RJ, Olson RS, Schmitt P, Meeker M, Moore JH. Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining. *Journal of Biomedical Informatics*. 2018;85:168–188. doi:10.1016/j.jbi.2018.07.015.
2. Urbanowicz RJ, Meeker M, Cava WL, Olson RS, Moore JH. Relief-Based Feature Selection: Introduction and Review. *Journal of Biomedical Informatics*. 2018;doi:10.1016/j.jbi.2018.07.014.
3. Robnik Šikonja M, Igor Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*. 2003;53:23 – 69.
4. Le TT, Dawkins BA, McKinney BA. Nearest-neighbor Projected-Distance Regression (NPDR) for detecting network interactions with adjustments for multiple tests and confounding. *Bioinformatics*. 2020;doi:10.1093/bioinformatics/btaa024.
5. Le TT, Urbanowicz RJ, Moore JH, McKinney BA. STatistical Inference Relief (STIR) feature selection. *Bioinformatics*. 2018; p. bty788. doi:10.1093/bioinformatics/bty788.
6. McKinney BA, White BC, Grill DE, Li PW, Kennedy RB, Poland GA, et al. ReliefSeq: a gene-wise adaptive-K nearest-neighbor feature selection tool for finding gene-gene interactions and main effects in mRNA-Seq gene expression data. *PloS one*. 2013;8(12):e81527.

7. Arabnejad M, Dawkins BA, Bush WS, White BC, Harkness AR, McKinney BA. Transition-transversion encoding and genetic relationship metric in ReliefF feature selection improves pathway enrichment in GWAS. *BioData Mining*. 2018;11(23).
8. Venkataraman A, Marek Kubicki, Carl-Fredrik Westin, Polina Golland. Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies. *Conf Comput Vis Pattern Recognit Workshops*. 2010; p. 63–70. doi:10.1109/CVPRW.2010.5543446.
9. Hay E, Petra Ritter, Nancy J Lobaugh, Anthony R McIntosh. Multiregional integration in the brain during resting-state fMRI activity. *PLOS Computational Biology*. 2017;doi:10.1371/journal.pcbi.1005410.
10. Sundermann B, Mona Olde lütke Beverborg, Bettina Pfleiderer. Toward literature-based feature selection for diagnostic classification: a meta-analysis of resting-state fMRI in depression. *Frontiers in Human Neuroscience*. 2014;doi:10.3389/fnhum.2014.00692.
11. Vergun S, Alok S Deshpande, Timothy B Meier, Jie Song, Dana L Tudorascu, Veena A Nair, et al. Characterizing functional connectivity differences in aging adults using machine learning on resting state fMRI data. *Frontiers in Computational Neuroscience*. 2013;doi:10.3389/fncom.2013.00038.
12. Le TT, Simmons WK, Misaki M, Bodurka J, White BC, Savitz J, et al. Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests. *Bioinformatics*. 2017;33(18):2906–2913.
13. Gotts SJ, Kyle Simmons W, Lydia A Milbury, Gregory L Wallace, Robert W Cox, Alex Martin. Fractionation of social brain circuits in autism spectrum disorders. *Brain*. 2012;135:2711–2725. doi:10.1093/brain/aws160.
14. Wasserman L. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, NY; 2004.
15. Miller I, Marylees Miller. *John E. Freund's Mathematical Statistics with Applications*. 7th ed. Yagan S, editor. Pearson Prentice Hall; 2004.
16. Brazma A, Jaak Vilo. Gene expression data analysis. *FEBS Letters*. 2000;480:17–24.
17. Wang D, Jin Gu. VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genomics Proteomics Bioinformatics*. 2018;16:320–331.
18. Gumbel EJ. The Distribution of the Range. *The Annals of Mathematical Statistics*. 1947;18(3):384–412.
19. Chatterjee S. *Superconcentration and Related Topics*. 1st ed. 1439-7382. Springer International Publishing; 2014.
20. Cramér H. *Mathematical Methods of Statistics*. vol. 1. Reprint, revised ed. Princeton University Press; 1999.
21. Li P, Maozu Guo, Chunyu Wang, Xiaoyan Liu, Quan Zou. An overview of SNP interactions in genome-wide association studies. *Briefings in Functional Genomics*. 2014;14(2):143–155. doi:10.1093/bfpg/elu036.

22. Lee MH, Christopher D Smyser, Joshua S Shimony. Resting state fMRI: A review of methods and clinical applications. *AJNR Am J Neuroradiol*. 2013;34(10):1866–1872. doi:10.3174/ajnr.A3263.
23. Alexander Dickie D, Susan D Shenkin, Devasuda Anblagan, Juyoung Lee, Manuel Blesa Cabez, David Rodriguez, et al. Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging. *Frontiers in Neuroinformatics*. 2017;doi:10.3389/fninf.2017.00001.