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

Bryan A. Dawkins[1], Trang T. Le[2] and Brett A. McKinney[1,3,*]
[1]Department of Mathematics, University of Tulsa, Tulsa, OK 74104, USA
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104
[3]Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104, USA.
∗Correspondence: 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 and interactions, are reflected in the expected moments of the distribution of pairwise distances. We derive general analytical expressions for the mean and variance of pairwise distances for $L_q$ metrics for 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, but empirically we also consider the effect of correlation. 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.