MULTIVARIATE INDEPENDENCE AND *K*-SAMPLE TESTING

by

Sambit Panda

A thesis submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Master of Science and Engineering

Baltimore, Maryland
May, 2020

Abstract

With the increase in the amount of data in many fields, a method to consistently and efficiently decipher relationships within high dimensional data sets is important. Because many modern datasets are multivariate, univariate tests are not applicable. While many multivariate independence tests have `R` packages available, the interfaces are inconsistent and most are not available in `Python`. We introduce `hyppo`, which includes many state of the art multivariate testing procedures. This thesis provides details for the implementations of each of the tests within a test `hyppo` as well as extensive power and runtime benchmarks on a suite of high-dimensional simulations previously used in different publications. The documentation and all releases for `hyppo` are available at https://hyppo.neurodata.io.

Primary Reader and Advisor: Joshua T. Vogelstein

Reader: Carey E. Priebe

Reader: Ronak Mehta

Acknowledgments

This work would not be possible without: the dedicated guidance from Dr. Joshua T. Vogelstein and Dr. Cencheng Shen, the support from Ronak Mehta, Eric W. Bridgeford, Satish Palaniappan, Junhao Xiong, the advice from Dr. Carey Priebe, and the Department of Biomedical Engineering at JHU.

Table of Contents

List of Tables

List of Figures

1 Introduction

Technological advancements have enabled the use of large amounts of data to represent important relationships in nearly every field. Examining and identifying relationships between sets of high-dimensional variables is critical to advance understanding and planning of future numerical and physical experiments. Independence and k-sample testing enables formally testing models to identify such differences.

Over the last century and a half, many different statistical tests have been developed to analyze such multivariate data sets. Early non-parametric tests were introduced in the 1940s and 1950s to test on distributions [1, 2]. In the 1970s and 1980s, nearest neighbor approaches were introduced that could operate on high dimensional and nonlinear sample data but required careful tuning of algorithm parameters [3, 4]. Recently, several statistics have been proposed that operate well on high-dimensional (potentially non-Euclidean) data, such as distance correlation [5–8] and Hilbert-Schmidt independence criterion [9–11], which are actually exactly equivalent in Sejdinovic *et al.* [12] and Shen & Vogelstein [13]. Heller, Heller and Gofrine proposed another nonparametric independence test with particularly high power in certain nonlinear relationships [14]. Multiscale Graph Correlation is a test that has demonstrated higher statistical power on many multivariate, nonlinear, and structured data when compared to other independence tests [15, 16], which combines and extends the nearest neighbors and energy statistics to detect underlying relationships. The test is statistically efficient, requiring about half or one-third of the number of samples to achieve the same statistical power

[17]. In addition, the test provides additional information about the data's geometry, allowing for more informed decision making of the underlying relationships in the data. For each of these tests, p-values can be calculated using a random permutation test [18–20]. These tests can be modified and extended to such applications as time-series testing [21].

To approach the problem of two-sample testing, Student's t-test [22] is traditionally used, while a few nonparametric alternatives have been proposed that operate well on multivariate, nonlinear data such as Energy [23], and maxmimal mean discrepency [24], and Heller Heller and Gorfine's test [14]. The two-sample testing problem can be generalized to the k-sample testing problem and here analysis of variance [25] or its multivariate analogue, multivariate analysis of variance [26], can be used, but these statistics either fail to, or operate poorly upon, multivariate and nonlinear data. In addition, both tests in particular suffer from fundamental assumptions that are not generally present in real data [27, 28]. There are a few nonparametric alternatives to analysis and multivariate analysis of variance, such as multivariate k-sample Heller Heller Gorfine [29], and distance components (DISCO) [30]. Recently, Shen *et al.* [31] has shown that nonparametric distance and kernel k-sample tests can be formulated by reducing the k-sample testing problem to the independence testing problem.

This thesis introduces `hyppo`, a comprehensive hypothesis package that provides various tests with high statistical power on multidimensional and nonlinear data. `hyppo` is a well-tested, multi-platform, `Python` 3 compatible library that allows users to conduct hypothesis tests on their data, and is also

flexible enough to allow developers to easily add in their own tests. hyppo is notable as it is one of the few packages in Python that enables such analysis and contains many uniquely power tests not present in other libraries. It also provides benchmarks for each of these tests by comparing power over many statistical models. The contribution of this thesis is therefore to provide: (1) an overview of notable independence tests implemented in hyppo, (2) benchmarks of the tests on a suite of diverse challenge problems, and (3) comparisons of the test statistics and wall times with similar R packages.

2 Preliminaries

2.1 Notation

Let $\mathbb{R}$ denote the real line $(-\infty, \infty)$. Let F_X, F_Y, and F_{XY} refer to the marginal and joint distributions of random variables X and Y respectively. Let x and y refer to the samples from F_X and F_Y and $\boldsymbol{x} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^{n \times q}$ refer to the matrix of these observations. That is, $\boldsymbol{x} = \{x_i \sim F_X \text{ where } x_i \in \mathbb{R}^p,\ i = 1, ..., n\}$ and $\boldsymbol{y} = \{y_i \sim F_Y \text{ where } y_i \in \mathbb{R}^q,\ i = 1, ..., n\}$. The trace of an $n \times n$ square matrix is the sum of the elements along the main diagonal; that is, the trace of $n \times n$ matrix $\boldsymbol{x}$ is $\text{tr}(\boldsymbol{x}) = \sum_{i=1}^{n} x_{ii}$.

2.2 Independence Tests

All independence tests can be generalized into the following form: given random variables X and Y, which are assumed to be from the joint distribution $F_{XY} = F_{X|Y} F_Y$, two variables are considered independent if and only if $F_{X|Y} F_Y = F_X F_Y$; that is, the joint distribution is equal to the product of the

marginals. This idea can be formulated as the following test:

$$H_0 : F_{XY} = F_X F_Y \qquad H_A : F_{XY} \neq F_X F_Y.$$

It turns out any dependency measure can be directly used to test equality of two or more distributions, i.e., two-sample or k-sample test, see [31].

2.2.1 Pearson's Product-Moment Correlation Coefficient (PEARSON), RV, and Canonical-Correlation Analysis (CCA)

PEARSON is a measure of the linear correlation between two univariate random variables [32]. Given sample data $\boldsymbol{x}$ and $\boldsymbol{y}$ where $p = q = 1$, the sample PEARSON correlation is

$$\text{PEARSON}_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{\widehat{\text{cov}}(\boldsymbol{x}, \boldsymbol{y})}{\hat{\sigma}_{\boldsymbol{x}} \hat{\sigma}_{\boldsymbol{y}}}, \tag{1}$$

where $\widehat{\text{cov}}(\boldsymbol{x}, \boldsymbol{y})$ is the sample covariance, $\hat{\sigma}_{\boldsymbol{x}}$ and $\hat{\sigma}_{\boldsymbol{y}}$ are the sample standard deviations of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively.

RV is a multivariate generalization of the squared PEARSON coefficient [33, 34]. The derivation is as follows: assuming each column in $\boldsymbol{x}$ and $\boldsymbol{y}$ are pre-centered to zero mean in each dimension, then the sample covariance matrix is $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \boldsymbol{x}^\mathsf{T} \boldsymbol{y}$, and the RV coefficient is

$$\text{RV}_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{\text{tr}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{yx}}\right)}{\text{tr}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{\boldsymbol{2}}\right) \text{tr}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{yy}}^{\boldsymbol{2}}\right)}. \tag{2}$$

Another similarly defined tool is CCA, which finds the linear combinations with respect to the dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}$ that maximize their correlation [35]. It seeks a vector $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^q$ to compute the first correlation coefficient

as

$$\max_{\boldsymbol{a}\in\mathbb{R}^n,\boldsymbol{b}\in\mathbb{R}^m} \frac{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_{xy}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \hat{\boldsymbol{\Sigma}}_{xx}\boldsymbol{a}}\sqrt{\boldsymbol{b}^\top \hat{\boldsymbol{\Sigma}}_{yy}\boldsymbol{b}}}. \tag{3}$$

One can keep on deriving the second and the third canonical correlation coefficients in a similar manner until the end, and CCA can also be generalized to more than two random variables ([36]). Therefore, CCA can be used to define a test statistic for dependence, and usually people take the first correlation coefficient or the sum of all correlation coefficients as the statistic.

2.2.2 Kendall (KENDALL) and Spearman (SPEARMAN)

KENDALL and SPEARMAN are rank-based correlation coefficients that are robust univariate test statistics [37, 38]. To formulate KENDALL, define (x_i, y_i) and (x_j, y_j) as concordant if the ranks agree: $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$. They are discordant if the ranks disagree: $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ and $y_i = y_j$, the pair is said to be tied. Let n_c and n_d be the number of concordant and discordant pairs respectively and $n_0 = n(n-1)/2$. In the case of no ties, the test statistic is defined as

$$\text{KENDALL}_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{n_c - n_d}{n_0}, \tag{4}$$

Further define $n_1 = \sum_i \frac{1}{2} t_i(t_i - 1)$, $n_2 = \sum_j \frac{1}{2} u_j(u_j - j)$, $t_i =$ number of tied values in the ith group of ties in the first quantity, and $u_j =$ number of tied values in the jth group of ties in the second quantity. In the case of ties, the statistic is calculated as in [39]

$$\text{KENDALL}_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)\,(n_0 - n_2)}}, \tag{5}$$

SPEARMAN can be thought of as closely related to PEARSON, whose statistic is listed in Equation 1. Suppose that rg_{x_i} and rg_{y_i} are the respective ranks of n raw scores x_i and y_i, ρ denotes the PEARSON coefficient but applied to rank variables, $\text{cov}(\text{rg}_{\boldsymbol{x}}, \text{rg}_{\boldsymbol{y}})$ denotes the covariance of the rank variables, and $\hat{\sigma}_{\text{rg}_x}$ and $\hat{\sigma}_{\text{rg}_y}$ denote the standard deviations of the rank variables. The statistic is

$$\text{SPEARMAN}_s(\boldsymbol{x}, \boldsymbol{y}) = \rho_{\text{rg}_x, \text{rg}_y} = \frac{\text{cov}(\text{rg}_{\boldsymbol{x}}, \text{rg}_{\boldsymbol{y}})}{\hat{\sigma}_{\text{rg}_x} \hat{\sigma}_{\text{rg}_y}}. \tag{6}$$

2.2.3 Heller, Heller, and Gorfine's (HHG)

	$d_{\boldsymbol{y}}(y_i, \cdot) \leq d_{\boldsymbol{y}}(y_i, y_j)$	$d_{\boldsymbol{y}}(y_i, \cdot) > d_{\boldsymbol{y}}(y_i, y_j)$	
$d_{\boldsymbol{x}}(x_i, \cdot) \leq d_{\boldsymbol{x}}(x_i, x_j)$	$A_{11}(i,j)$	$A_{12}(i,j)$	$A_{1\cdot}(i,j)$
$d_{\boldsymbol{x}}(x_i, \cdot) > d_{\boldsymbol{x}}(x_i, x_j)$	$A_{21}(i,j)$	$A_{22}(i,j)$	$A_{2\cdot}(i,j)$
	$A_{\cdot 1}(i,j)$	$A_{\cdot 2}(i,j)$	$n-2$

Table 1: The cross-cross classification table used to calculate the Pearson's chi squared test statistic involved in the HHG test statistic calculation.

HHG is a consistent multivariate test of associations based on the rank of the distances [14]. For every sample point $j \neq i$, denote a point in the joint sample space as (x_j, y_j). Let $d_{\boldsymbol{x}}(x_i, x_j)$ be equivalent to the norm distance between samples x_i and x_j and $d_{\boldsymbol{y}}(y_i, y_j)$ is similarly defined. The indicator function is denoted by $\mathbb{I}\{\cdot\}$. The cross-classification between these two random variables can be formulated as in Table 1, where

$$A_{11} = \sum_{k=1, k\neq i,j}^{n} \mathbb{I}\left\{d_{\boldsymbol{x}}(x_i, x_k) \leq d_{\boldsymbol{x}}(x_i, x_j)\right\} \mathbb{I}\left\{d_{\boldsymbol{y}}(y_i, y_k) \leq d_{\boldsymbol{y}}(y_i, y_j)\right\},$$

and A_{12}, A_{21}, and A_{22} are defined similarly. $A_{\cdot 1}$, $A_{\cdot 2}$, $A_{1\cdot}$, and $A_{2\cdot}$ are the sums of the column and row respectively. Once this table is generated, the

Pearson's chi square test statistic can be calculated using

$$S(i,j) = \frac{(n-2)(A_{12}A_{21} - A_{11}A_{22})^2}{A_{1\cdot}A_{2\cdot}A_{\cdot 1}A_{\cdot 2}}.$$

From here, the HHG test statistic is simply

$$\text{HHG}_n = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} S(i,j). \tag{7}$$

2.2.4 DCORR

DCORR is a powerful test to determine linear and nonlinear associations between two random variables or vectors in arbitrary dimensions. The test statistic can be determined as follows: let $\boldsymbol{D}^{\boldsymbol{x}}$ be the $n \times n$ distance matrix of $\boldsymbol{x}$ and $\boldsymbol{D}^{\boldsymbol{y}}$ be the $n \times n$ distance matrix of $\boldsymbol{y}$. Let $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{J}$ denote the $n \times n$ centering matrix where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{J}$ is the matrix of ones. The distance covariance (DCOV') and distance correlation (DCORR') can then be defined as [7],

$$\text{DCOV}'_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n^2}\text{tr}(\boldsymbol{H}\boldsymbol{D}^{\boldsymbol{x}}\boldsymbol{H}\boldsymbol{H}\boldsymbol{D}^{\boldsymbol{y}}\boldsymbol{H}). \tag{8}$$

$$\text{DCORR}'_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{\text{DCOV}'_n(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\text{DCOV}'_n(\boldsymbol{x}, \boldsymbol{x}) \cdot \text{DCOV}'_n(\boldsymbol{y}, \boldsymbol{y})}} \in [-1, 1]. \tag{9}$$

The statistics presented in equations (8) and (9) are biased; fortunately, unbiased distance correlation test statistics have also been developed [40]. Define another modified matrix $\boldsymbol{C}^{\boldsymbol{x}}$ such that,

$$\boldsymbol{C}_{ij}{}^{\boldsymbol{x}} = \begin{cases} \boldsymbol{D}^{\boldsymbol{x}}_{ij} - \frac{1}{n-2}\sum_{t=1}^{n} \boldsymbol{D}^{\boldsymbol{x}}_{it} - \frac{1}{n-2}\sum_{s=1}^{n} \boldsymbol{D}^{\boldsymbol{x}}_{sj} + \frac{1}{(n-1)(n-2)}\sum_{s,t=1}^{n} \boldsymbol{D}^{\boldsymbol{x}}_{st} & i \neq j \\ 0 & \text{otherwise} \end{cases},$$

and define $\boldsymbol{C}^{\boldsymbol{y}}$ similarly. Then, the unbiased distance covariance (DCOV) and unbiased distance correlation (DCORR) is [40],

$$\text{DCOV}_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n(n-3)} \text{tr}(\boldsymbol{C}^{\boldsymbol{x}} \boldsymbol{C}^{\boldsymbol{y}}). \tag{10}$$

$$\text{DCORR}_n(\boldsymbol{x}, \boldsymbol{y}) = \frac{\text{DCOV}_n(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\text{DCOV}_n(\boldsymbol{x}, \boldsymbol{x}) \cdot \text{DCOV}_n(\boldsymbol{y}, \boldsymbol{y})}} \in [-1, 1]. \tag{11}$$

Since the statistics presented in equations (10) and (11) provide similar empirical results to the biased statistics [17], from now on any reference to distance correlation will refer to the unbiased distance correlation. In fact, this formulation of k-sample DCORR is exactly equivalent to energy distance [31].

2.2.5 HSIC

Hilbert-Schmidt independence criterion (HSIC) is a closely related test that exchanges distance matrices $\boldsymbol{D}^{\boldsymbol{x}}$ and $\boldsymbol{D}^{\boldsymbol{y}}$ for kernel similarity matrices $\boldsymbol{K}^{\boldsymbol{x}}$ and $\boldsymbol{K}^{\boldsymbol{y}}$. In fact, they are exactly equivalent in the sense that every valid kernel has a corresponding valid semimetric to ensure their equivalence, and vice versa [12, 13]. In other words, every DCORR test is also an HSIC test and vice versa. Nonetheless, implementations of DCORR and HSIC use different metrics by default: DCORR uses a Euclidean distance while HSIC uses a Gaussian kernel similarity.

2.2.6 MGC

Building upon the ideas of DCORR, HSIC, and k-nearest neighbors, MGC preserves the consistency property while typically working better in multivariate

and non-monotonic relationships [17]. The MGC test statistic is computed as follows:

1. Two distance matrices $\boldsymbol{D}^{\boldsymbol{x}}$ and $\boldsymbol{D}^{\boldsymbol{y}}$ are computed, and modified to be mean zero column-wise. This results in two $n \times n$ distance matrices $\boldsymbol{C}^{\boldsymbol{x}}$ and $\boldsymbol{C}^{\boldsymbol{y}}$ (the centering and unbiased modification is slightly different from the unbiased modification in the previous section, see [15] for more details).

2. For all values k and l from $1, \ldots, n$,

   (a) The k-nearest neighbor and l-nearest neighbor graphs are calculated for each property. Here, $\boldsymbol{G}_k(i,j)$ has value 1 for the k smallest values of the i-th row of $\boldsymbol{D}^{\boldsymbol{x}}$ and $\boldsymbol{H}_l(i,j)$ has value 1 the l smallest values of the i-th row of $\boldsymbol{D}^{\boldsymbol{y}}$. All other values in both matrices is 0.

   (b) The local correlations are summed and normalized using the following statistic:

$$c^{kl} = \frac{\sum_{ij} \boldsymbol{D}^{\boldsymbol{x}}(i,j)\boldsymbol{G}_k(i,j)\boldsymbol{D}^{\boldsymbol{y}}(i,j)\boldsymbol{H}_l(i,j)}{\sqrt{(\boldsymbol{D}^{\boldsymbol{x}}(i,j))^2 \boldsymbol{G}_k(i,j)} \cdot \sqrt{(\boldsymbol{D}^{\boldsymbol{y}}(i,j))^2 \boldsymbol{H}_l(i,j)}},$$

3. The MGC test statistic is the smoothed optimal local correlation of $\{c^{kl}\}$. Denote the smoothing operation as $R(\cdot)$ (which essentially set all isolated large correlations as 0 and connected large correlations same as before, see [15]), MGC is

$$\text{MGC}_n(\boldsymbol{x}, \boldsymbol{y}) = \max_{(k,l)} R(c^{kl}(\boldsymbol{x}_n, \boldsymbol{y}_n)). \tag{12}$$

2.3 Two-sample and *k*-sample Tests

Consider the two-sample problem: we obtain two datasets: $u_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$ and $v_j \in \mathbb{R}^p$ for $j = 1, \ldots, m$. Assume that each u_i is sampled independently and identically (i.i.d.) from F_U and that each v_j is sampled i.i.d. from F_V (and also that each u_i and each v_j is independent from one another). The two-sample testing problem tests whether the two datasets were sampled from the same distribution, that is,

$$H_0 : F_U = F_V, \qquad H_A : F_U \neq F_V. \tag{13}$$

Eq. (13) can also be generalized to k samples: let $x_{j_i} \in \mathbb{R}^p$ for $j = 1, \ldots, k$ and $i = 1, \ldots, n_j$ be k datasets that are sampled i.i.d. from $F_1, \ldots, F_k$ and independently from one another. Then,

$$H_0 : F_1 = F_2 = \cdots = F_k, \qquad H_A : \exists\, j \neq j' \text{ s.t. } F_j \neq F_{j'} \tag{14}$$

2.3.1 HOTELLING

HOTELLING is a generalization of Student's t-test in arbitrary dimension [41]. Consider input samples $\boldsymbol{u}_i \overset{iid}{\sim} F_U$ for $i \in \{1, \ldots, n\}$ and $\boldsymbol{v}_i \overset{iid}{\sim} F_V$ for $i \in \{1, \ldots, m\}$. Let $\bar{\boldsymbol{u}}$ refer to the columnwise means of $\boldsymbol{u}$; that is, $\bar{\boldsymbol{u}} = (1/n) \sum_{i=1}^{n} \boldsymbol{u}_i$ and let $\bar{\boldsymbol{v}}$ be the same for $\boldsymbol{v}$. Calculate sample covariance matrices $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{uv}} = \boldsymbol{u}^T \boldsymbol{v}$ and sample variance matrices $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{uu}} = \boldsymbol{u}^T \boldsymbol{u}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{vv}} = \boldsymbol{v}^T \boldsymbol{v}$. Denote pooled covariance matrix $\hat{\boldsymbol{\Sigma}}$ as

$$\hat{\boldsymbol{\Sigma}} = \frac{(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{uu}} + (m-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{vv}}}{n+m-2}$$

Then,

$$\textsc{Hotelling}_{n,m}(\boldsymbol{u}, \boldsymbol{v}) = \frac{nm}{n+m}(\bar{\boldsymbol{u}} - \bar{\boldsymbol{v}})^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{u}} - \bar{\boldsymbol{v}}) \tag{15}$$

Of course, since it is a multivariate generalization of Student's t-tests, it suffers from some of the same assumptions as Student's t-tests. That is, the validity of HOTELLING depends on the assumption that random variables are normally distributed within each group, and each with the same covariance matrix. Distributions of input data are generally not known and cannot always be reasonably modeled as Gaussian [42, 43], and having the same covariance across groups is also generally not true of real data.

2.3.2 MANOVA

MANOVA is a procedures for comparing more than two multivariate samples [27, 44]. It can be thought as a multivariate generalization of the univariate ANOVA [27] using covariance matrices rather than the scalar variances. As in Rencher [45]: consider input samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ that have the same dimensionality p. Each $\boldsymbol{x}_i$, where $i \in \{1, \ldots, k\}$ is assumed to be sampled from a multivariate distribution $\mathbb{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ and so each sample is assumed to have the same covariance matrix $\boldsymbol{\Sigma}$. The model for each p-dimensional vector of each $\boldsymbol{x}_i$ is defined as follows: for $j \in \{1, \ldots, n_i\}$,

$$\boldsymbol{x}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij}.$$

In MANOVA, we are testing if the mean vectors of each of the k-samples is the same. That is, the null and alternate hypotheses are,

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_k, \qquad H_A : \exists\, j \neq j' \text{ s.t. } \boldsymbol{\mu}_j \neq \boldsymbol{\mu}_{j'}$$

Let $\bar{\boldsymbol{x}}_{i\cdot}$ refer to the columnwise means of $\boldsymbol{x}_i$; that is, $\bar{\boldsymbol{x}}_{i\cdot} = (1/n_i)\sum_{j=1}^{n_i} \boldsymbol{x}_{ij}$. The pooled sample covariance of each group, $\boldsymbol{W}$, is

$$\boldsymbol{W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{i\cdot})(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{i\cdot})^{\mathsf{T}}. \tag{16}$$

Next, define $\boldsymbol{B}$ as the sample covariance matrix of the means. If $N = \sum_{i=1}^{k} n_i$ and the grand mean is $\bar{\boldsymbol{x}}_{\cdot\cdot} = (1/N)\sum_{i=1}^{k}\sum_{j=1}^{n} \boldsymbol{x}_{ij}$,

$$\boldsymbol{B} = \sum_{i=1}^{k} n_i(\bar{\boldsymbol{x}}_{i\cdot} - \bar{\boldsymbol{x}}_{\cdot\cdot})(\bar{\boldsymbol{x}}_{i\cdot} - \bar{\boldsymbol{x}}_{\cdot\cdot})^{\mathsf{T}}. \tag{17}$$

Some of the most common statistics used when performing MANOVA include the Wilks' Lambda, the Lawley-Hotelling trace, Roy's greatest root, and Pillai-Bartlett trace (PBT) [46–48] (PBT is recognized to be the best of these as it is the most conservative [27, 49]) and Olson [50] has shown that there is minimal differences in statistical power among these statistics. Let $\lambda_1, \lambda_2, \ldots, \lambda_s$ refer to the eigenvalues of $(\boldsymbol{B} + \boldsymbol{W})^{-1}\boldsymbol{B}$. Here $s = \min(\nu_{\boldsymbol{B}}, p)$ is the minimum between the degrees of freedom of $\boldsymbol{B}$, $\nu_{\boldsymbol{B}}$ and p. So, the PBT MANOVA test statistic can be written as [44],

$$\textsc{Manova}_{n_1,\ldots,n_k}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} = \operatorname{tr}\Big((\boldsymbol{B} + \boldsymbol{W})^{-1}\boldsymbol{B}\Big). \tag{18}$$

MANOVA is closely related to HOTELLING, and as such, it suffers from the same assumptions that HOTELLING does.

2.3.3 k-sample Tests as Independence Tests

k-sample tests can be implemented as independence tests as follows: consider $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ as matrices of size $n_1 \times p, \ldots, n_k \times p$, where p refers to the number of

dimensions and n_i refers to the number of samples of $\boldsymbol{u}_i$. Letting $N = \sum_{i=1}^{k} n_i$, define new data matrices $\boldsymbol{x}$ and $\boldsymbol{y}$ such that,

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_k \end{bmatrix} \in \mathbb{R}^{N \times p},$$

$$\boldsymbol{y} = \begin{bmatrix} \mathbf{1}_{n_1 \times 1} & \mathbf{0}_{n_1 \times 1} & \dots & \mathbf{0}_{n_1 \times 1} \\ \mathbf{0}_{n_2 \times 1} & \mathbf{1}_{n_2 \times 1} & \dots & \mathbf{0}_{n_2 \times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_k \times 1} & \mathbf{0}_{n_k \times 1} & \dots & \mathbf{1}_{n_k \times 1} \end{bmatrix} \in \mathbb{R}^{N \times k}.$$

Additionally, in the two-sample case,

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} \in \mathbb{R}^{N \times p},$$

$$\boldsymbol{y} = \begin{bmatrix} \mathbf{0}_{n_1 \times 1} \\ \mathbf{1}_{n_2 \times 1} \end{bmatrix} \in \mathbb{R}^{N}.$$

That is, $\boldsymbol{x}$ can be thought of as the data matrix (contains all the concatenated data) while $\boldsymbol{y}$ can be thought of as the label matrix (labels $\boldsymbol{x}$ from whichever original input the data came from). Therefore, $\boldsymbol{x}$ and $\boldsymbol{y}$ are now paired data matrices, and thus dependence of $\boldsymbol{x}$ on $\boldsymbol{y}$ indicates that the labels are informative; in other words, that $\boldsymbol{u}$ and $\boldsymbol{v}$ have been sampled from different distributions. The implication of this idea is that any independence test can be used to implement a k-sample test [31]. Using this method, ENERGY is equivalent to two-sample DCORR and two-sample HSIC is equivalent to maximum mean discrepancy (MMD) exactly [31].

2.4 Permutation Tests

For many early independence tests, such as Pearson's, analytical p-values are available. When such analytic approximations are unknown, permutation tests permute either of the input data matrices $\boldsymbol{x}$ or $\boldsymbol{y}$, and calculate test statistics for each permutation. Doing so many times approximates the null distribution from which the observed test statistic can be compared to generate a p-value [51, 52].

In the case of nonparametric tests, permutations can be used to exactly calculate the p-value since calculations are not dependent upon a reference distribution [53]. However, in the case of large amounts of data, calculating every permutation is impractical and often computationally expensive. A finite number of permutations typically approximates the true null distribution quite well with a minimal additional computational cost [19, 53]. All tests that are used in section 3 use this permutation method to approximate a p-value.

2.5 hyppo

`hyppo` is an open source `Python` package that implements all the aforementioned tests in an easy-to-use and extensible framework. Links to source code, documentation, and tutorials can be found here: `https://hyppo.neurodata.io`. The modules of `hyppo` are: `independence`, `ksample`, `time_series`, and `sims`. Each module contains a `base.py` which contains the base abstract class for each module and a private `_utils.py` files that contains an input checking class and other relevant functions used by multiple classes in the module. Also, a single `Python` file contains a class corresponding to each independence

or k-sample test and any other private functions unique to statistic calculation. The p-value calculation uses a permutation test by default in most modules, this is overridden when analytical p-values are available for large sample sizes. Each test within hyppo contains a `.test` method which the user runs that returns at least a statistic and p-value in all cases. `sims` contains a benchmarks suite of 20 simulations to test statistical power of each of the tests in hyppo.

Code is released under the Apache v2.0 license on GitHub with releases available via PyPi. Documentation also details some background behind calculating each test statistic and links to relevant papers about each algorithm. Implementation follows PEP8 and has a high level of test coverage (> 85%). Development undergoes continuous integration and testing on Windows, Ubuntu Linux, and Mac OS X for Python 3.5+ and can be found on GitHub at hyppo.

2.6 Evaluating Implementations

To effectively evaluate implementations of each of the included independence tests, a number of jupyter notebooks have been written to evaluate speed, correctness, and power. The testing power for a given level of α (Type 1 error level) test is equal to the probability of correctly rejecting the null hypothesis when the alternative is true. For a test to be consistent, statistical power must converge to 1 as the sample size increases to ∞. To this end, a benchmark of 20 different distributions, as developed previously for independence testing, including polynomial (linear, quadratic, cubic), trigonometric (sinusoidal,

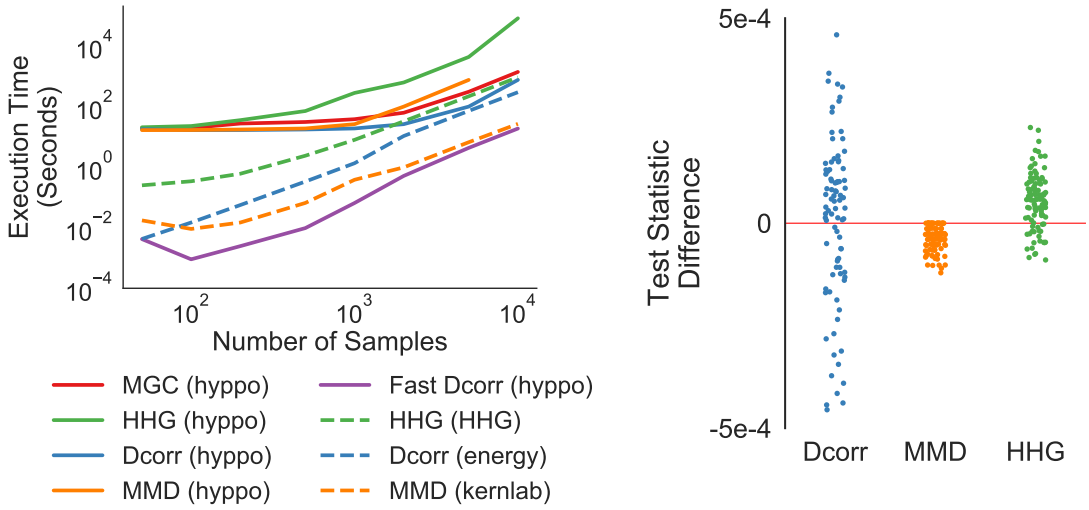

Figure 1: Benchmarks of `hyppo` implementations against corresponding `R` implementations. Average wall times (over 3 repetitions) (left) are shown for DCORR in `energy` and `kernlab` as compared against `hyppo` implementations of MGC, DCORR, and FAST DCORR. Test statistic comparisons (right) between DCORR, MMD, and HHG in `hyppo` are compared against their respective reference `R` implementations. Test statistics are nearly identical for each implementation.

circular, ellipsoidal, spiral), geometric (square, diamond, W-shaped), and other relationships [6, 7, 14, 17, 54, 55]. These distributions have been incorporated in the the `sims` module in `hyppo` with modifications made to test for k-samples.

3 Results

3.1 hyppo Benchmarks

3.1.1 Wall Times

Figure 1a shows the computational efficiency of `hyppo`'s implementations against existing implementations in commonly used `R` packages—specifically `energy` [56], `kernlab` [57], and `HHG` [58]. When comparing performance, wall

times are averages of p-value computations (1000 replications when permutation tests are used) 3 trials calculated on a univariate noisy linear simulation with number of samples increasing from 50 to 10,000. All computations were performed on an Ubuntu 18.04.3 LTS system with access to 96 cores. When sample sizes are above a few hundred, all algorithms achieve approximately quadratic times, with different slopes. HHG was the slowest as expected, though had comparable speeds to the other algorithms at low sample sizes. MGC and DCORR are next, and still only requires tens of minutes to run when sample sizes are around 10,000. At low sample sizes, the `energy` package's DCORR is faster than `kernlab`'s implementation of MMD (DCORR is equivalent to MMD for all finite sample sizes [13]) even at a sample size of 10,000. `hyppo`'s FAST DCORR is the fastest, even though both `energy` and `kernlab` both use highly optimized C++ versions.

3.1.2 Implementation Validation

Next, we verify that `hyppo`'s test statistics are equivalent to existing `R` implementations of the tests. Specifically, `hyppo`'s implementations were compared to: DCORR from the `energy` package [56]. MMD from the `kernlab` package [57], and HHG from the `HHG` package [58]. The evaluation uses a spiral simulation with 1000 samples and 2 dimensions for each test and compares test statistics over 20 repetitions. Figure 1b shows the difference between the `hyppo` implementation of the independence test and the respective `R` package implementation of the independence test. Although a slight numerical bias exists in the case of MMD due to a transformation from Shen & Vogelstein [13] and Shen *et al.* [31], test statistics are nearly equivalent for each

implementation.

3.2 Independence Power

Power curves were created for increasing sample size (Figure 2) and increasing dimension (Figure 3) for 20 different simulation settings, and were implemented from the equations in Appendix A. In all cases, $\alpha = 0.05$.

Figure 2 shows the effect of increasing sample size for each simulations has on the statistic power for each independence test. Number of samples ranged from 5 to 100 and the fast tests started at 20 samples since that is the minimum size that the tests need to operate. For each test and simulation, for all universally consistent tests (including MGC, DCORR, HSIC, PEARSON, SPEARMAN, CCA, HHG, KENDALL, and RV) it is expected that the statistical power will converge to one as the number of samples increased except in the case of simulation 20, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent. Better independence tests converge to one faster. 100 samples was chosen as the maximum because for most simulations, the power approached or was approaching one. In the first 14 settings, MGC performs as well or better than all the other tests and HHG perform best in the last four dependence settings. RV, CCA, KENDALL, SPEARMAN, and PEARSON perform poorly in all but the five monotonic settings (in the top row). All tests are valid, as shown in the last panel.

Figure 3 shows the effect of increasing dimensions on statistical power for sample size fixed at $n = 100$. The maximum number of dimensions varied for each simulation settings; it was higher for relatively simple relationships, and lower for relatively complex relationships. It is expected that power would

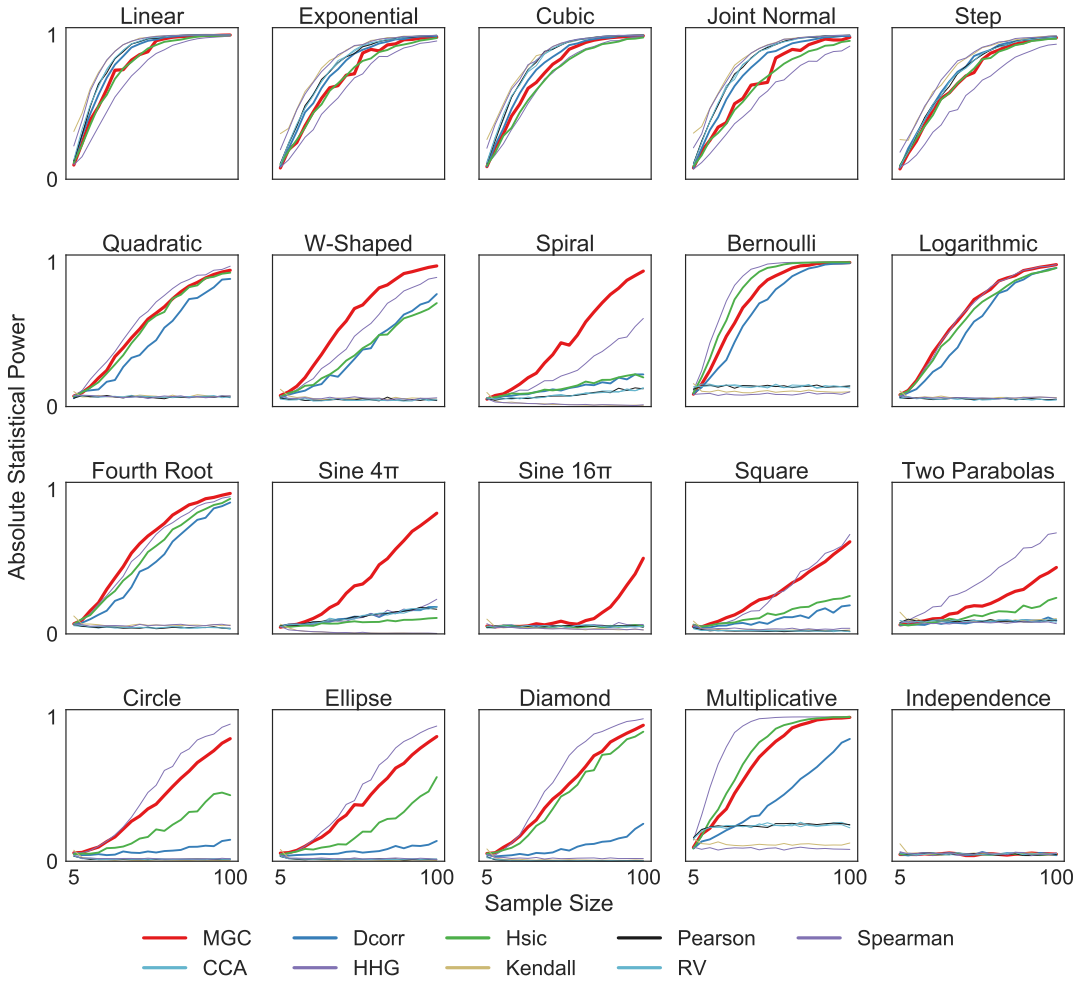

Figure 2: Power vs. sample size curves for each of 20 simulations with one dimension for both $\boldsymbol{x}$ and $\boldsymbol{y}$ ($n = 5$ trials). The fast tests require a sample size of at least 20 to calculate a reliable estimate of the p-value. MGC tends to perform better or the best among all the independence tests. Under the Multimodal Independence simulation, all tests achieve a power equal to α, as expected.

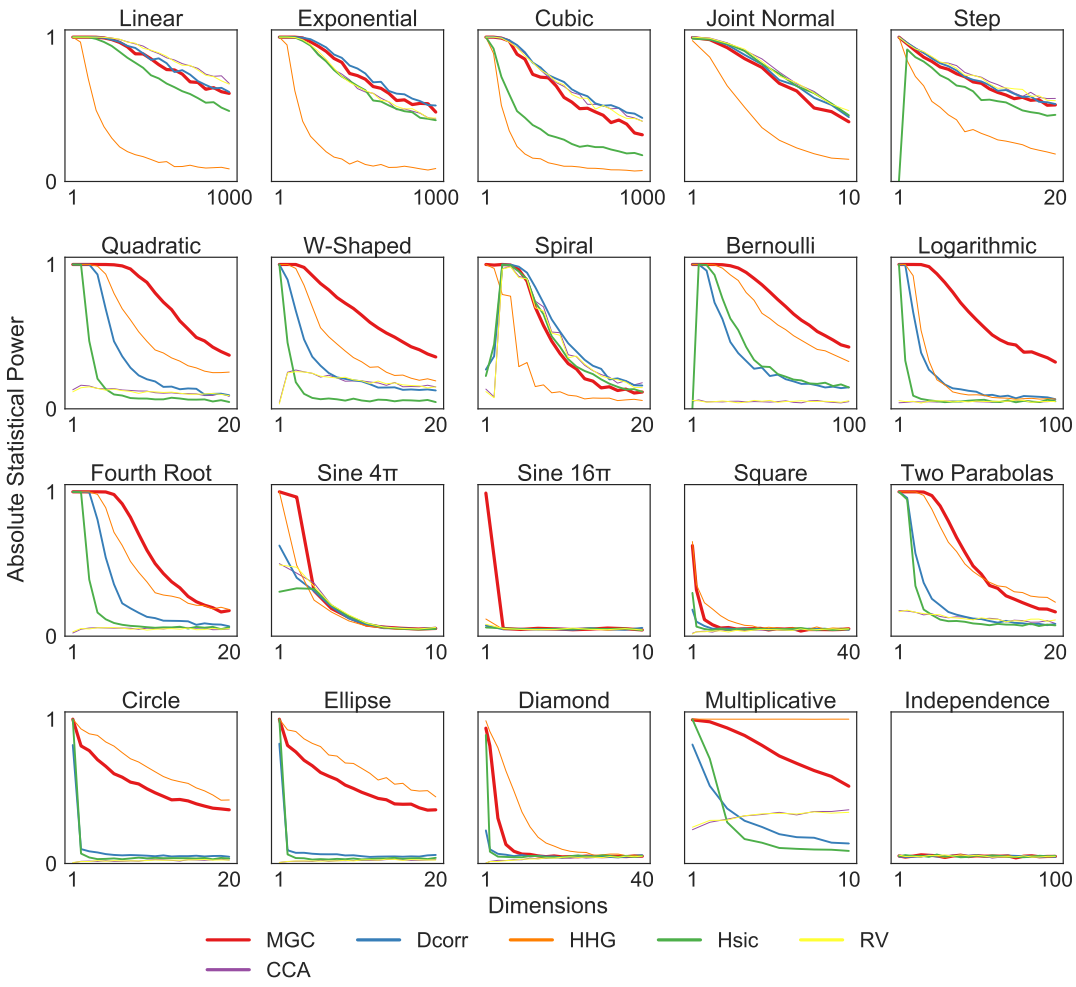

Figure 3: Power vs. dimension curves for each of 20 simulations using 100 samples for each different dimension ($n = 5$ trials). The results are qualitatively similar to those from Figure 2.

decrease toward 0 as the number of dimensions increase because adding dimensions increases the complexity of the simulation relationships and thus leading to more unreliable p-value estimate. Better tests converge to 0 more slowly. Under these conditions, as before, MGC tends to be either the best or near the best for most settings, except the last five, where HHG achieve slightly higher power than MGC for nearly all dimensions. RV and CCA all perform poorly for any of the non-monotonic settings (all but the top row).

3.3 *k*-sample Power

3.3.1 Gaussian Simulations

Consider the simplest possible three-sample tests, where in each case, all three samples are Gaussian with identity covariance matrix (I):

1. **None Different** All three groups are Gaussian with the same mean: $\mu = (0, 0)$.

2. **One Different** Two of the Gaussians have the same mean while the third has a different mean, thus, $\mu = (0, 0)$ for two of the Gaussians and $\mu = (0, \epsilon)$ for the third Gaussian.

3. **All Different** The three means form an equilateral triangle with center $(0, 0)$ and radius ϵ, thus,
   $\mu_1 = (0, \sqrt{3}/3 \times \epsilon)$, $\mu_2 = (-\epsilon/2, -\sqrt{3}/6 \times \epsilon)$, and $\mu_3 = (\epsilon/2, -\sqrt{3}/6 \times \epsilon)$.

Figure 4 shows (top) scatter plots and (bottom) statistical power for each

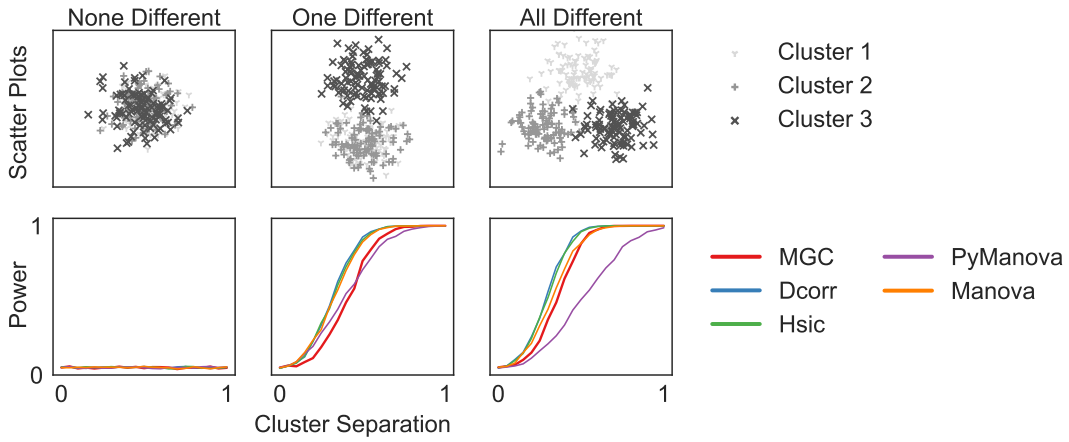

Figure 4: Comparing parametric to non-parameteric tests on three different parametric settings. Three two-dimensional Gaussians were generated for three different cases with 100 samples each (see Section A for details). The top row shows a scatter plot of each simulation for a given cluster separation, and the bottom row shows the power curves for each simulation as cluster separation increases (averaged over 5 repetitions). All methods are valid because power is $\leq \alpha$ (left). Shockingly, even on in a Gaussian settings in which one would expect MANOVA to be best, DCORR and HSIC perform as well (middle) or *better than* MANOVA (right). *k*-sample MGC adds a little variance to DCORR and HSIC which reduces it power relatively in these settings by a little. PYMANOVA performs poorly.

of the three cases, where ϵ is increased from 0 to 1. **None Different** demonstrates that each test controls type I error properly. Since there is no difference in distribution, all tests are expected to have power equal to α (0.05 in this case). **One Different** shows that as one distribution separates from the others, *k*-sample DCORR and *k*-sample HSIC perform similarly to MANOVA while slightly outperforming both PYMANOVA and *k*-sample MGC. In **All Different**, both *k*-sample DCORR and *k*-sample HSIC slightly outperform MANOVA, which performs similarly to *k*-sample MGC, and PYMANOVA performs particularly poorly. These results suggests that even at a simulation setting where the MANOVA test is expected to perform the best (linear simulation setting, all

distributions Gaussian, all distributions same covariance), nonparametric k-sample tests can perform as well, and even a little better! We do not consider PyMANOVA further.

3.3.2 A Benchmark Suite of 20 Affine Transformations

We consider a benchmark suite of 20 different distributions as developed previously for independence testing, including polynomial (linear, quadratic, cubic), trigonometric (sinusoidal, circular, ellipsoidal, spiral), geometric (square, diamond, W-shaped), and other relationships [6, 7, 14, 17, 54, 55, 59] with math and visualization shown in section A. In each case, we sample n times from one of these 20 different distributions, and then apply an affine transformation to the distribution, and sample n times again (so, in the following, $n = m$ for all simulations). In each case, the noise distribution is determined as described in Vogelstein *et al.* [17]. Figure 8 shows an example where we applied a 60 degree rotation to each distribution to obtain two samples with no noise. The following three figures show power curves for each of the 20 settings. The bottom right panel illustrates the power under the null, which must be less than or equal to α to be a valid test.

Figure 5 evaluates the tests for varying sample size in two-sample tests where both x and y are two-dimensional, and F_Y is rotated 90 degrees relative to F_X. The y-axis shows the power of each test relative to k-sample MGC's power (red line), meaning that if a test achieves higher power than MGC its curve is above the red line, and otherwise its curve is below the red line. In this setting, k-sample MGC performs as well or better than all other k-sample tests

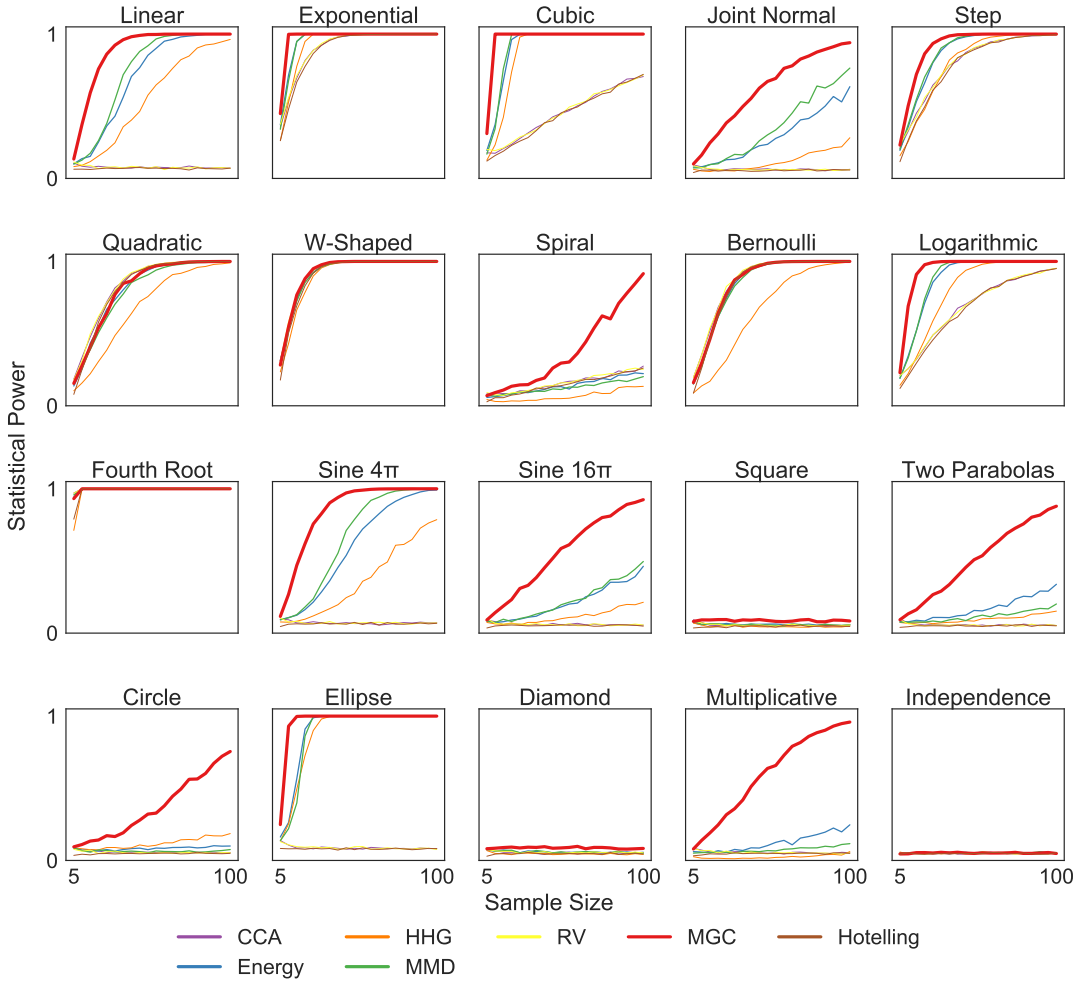

Figure 5: Power versus sample size curves for each of 20 two-sample simulations for a fixed angle (90 degrees), where both x and y are two-dimensional (averaged over 5 repetitions). Power curves are plotted relative to k-sample MGC: those above the red line outperform k-sample MGC and those under the red line perform worse than k-sample MGC. k-sample MGC empirically dominates all other tests, meaning it always achieves as high or higher statistical power for all simulations and sample sizes.

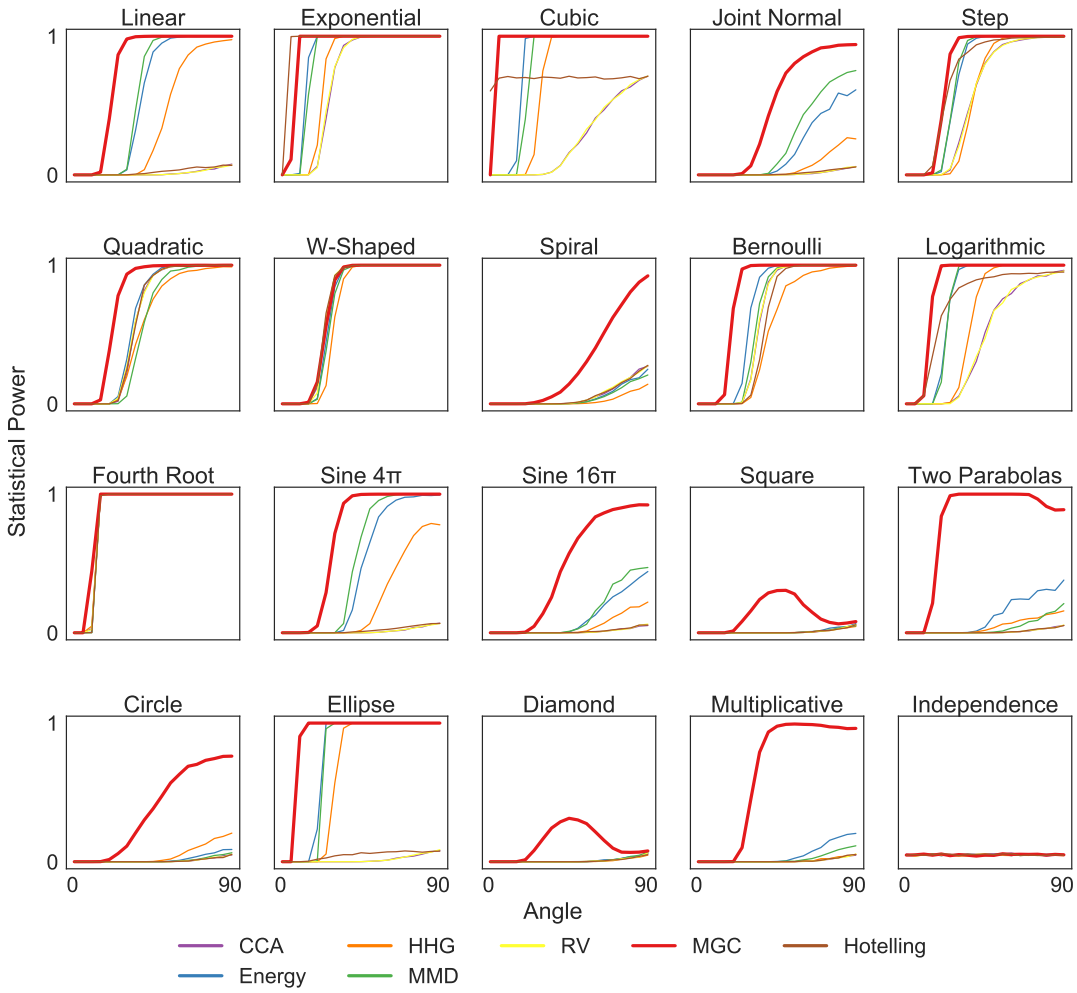

Figure 6: Power versus angle for 20 two-sample tests with fixed sample size (100 samples) in two dimensions (averaged over 5 repetitions). *k*-sample MGC empirically dominates the other tests in nearly all of the simulation settings. MANOVA performs slightly better than MGC for certain sample sizes in both the exponential and cubic simulations, probably because those settings closely approximate the setting MANOVA was designed for.

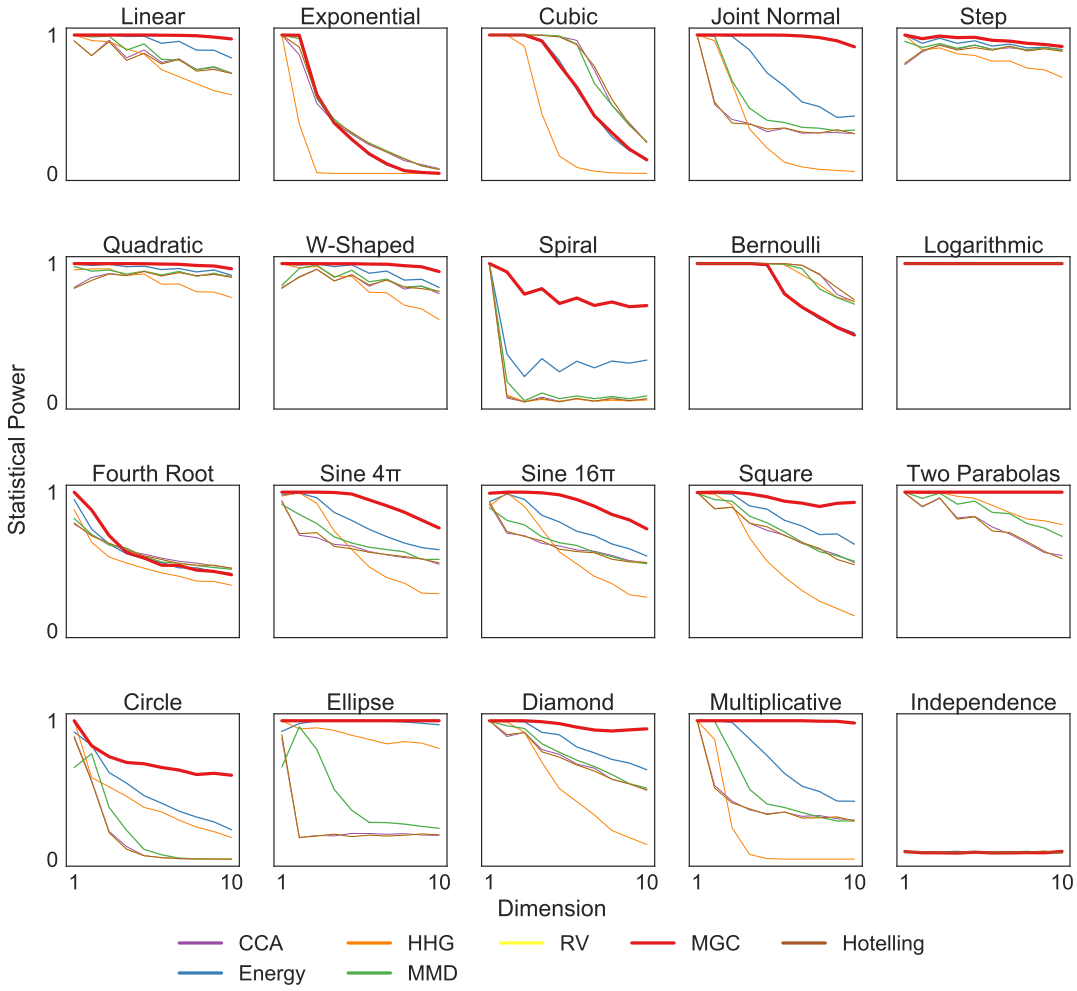

Figure 7: Power versus dimension for 20 two-sample tests with fixed sample size (100 samples) and angle (90 degrees) in two-dimensions (averaged over 5 repetitions). *k*-sample MGC empirically dominates the other tests in nearly all of the simulation settings. MANOVA performs slightly better than MGC for certain sample sizes in both the cubic and Bernoulli simulations, probably because those settings closely approximate the setting MANOVA was designed for.

in all simulation settings and sample sizes while properly controlling Type I error. MANOVA performs similarly to k-sample CCA and k-sample RV. Note that k-sample MGC outperforms MANOVA even in the linear setting.

Figure 6 shows the same 20 settings, exact the sample size fixed at $n = 100$ the rotation angle for F_Y is varied from 0° to 90°. As with Figure 5, power was plotted relative to k-sample MGC. In this setting, for nearly all angles and simulation settings, k-sample MGC achieved the same or higher power as every other test in nearly all settings. Here, however, MANOVA briefly outperforms k-sample MGC in two simulation settings (exponential and cubic). Visually inspecting these settings indicates that these settings are approximately Gaussian, where we previously demonstrated MANOVA can achieve higher power than MGC for certain parameter settings and sample size combinations.

Figure 7 shows the power as the number of dimensions is increasing, while the sample size and angle are fixed at 100 and 90 degrees, respectively. k-sample MGC outperformed every test in nearly all settings again. MANOVA performed better in the cubic and Bernoulli simulations; visual inspection indicates that these two settings can reasonably be approximated by Gaussians.

4 Conclusion

We have presented a number of known and novel independence tests that we have incorporated into a Python package `hyppo`. `hyppo` is an extensive and extensible open-source `Python` package for multivariate hypothesis testing. Incorporated within this package are a number of k-sample-tests based on a

trivial modification of existing independence tests [31]. It is an easy to use tool for anyone familiar with machine learning and `Python`. Applications of this work are far reaching within many fields from machine learning and artificial intelligence to general chemistry and biology. Hypothesis testing is a fundamental necessity in data analysis and having one in `Python`, which is a very commonly used programming language, is important. As `hyppo` continues to grow and add functionality, it will enhance tools scientists use when determining relationships within their investigations.

References

1. Hoeffding, W. A non-parametric test of independence. *The Annals of Mathematical Statistics,* 546–557 (1948).
2. Rényi, A. On measures of dependence. *Acta mathematica hungarica* **10,** 441–451 (1959).
3. Friedman, J. H. & Rafsky, L. C. Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics* **11,** 377–391 (1983).
4. Schilling, M. F. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association* **81,** 799–806 (1986).
5. Székely, G. J. & Rizzo, M. L. Brownian distance covariance. *The Annals of Applied Statistics* **3,** 1236–1265 (2009).
6. Székely, G. J. & Rizzo, M. L. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis* **117,** 193–213 (2013).
7. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35,** 2769–2794 (2007).
8. Lyons, R. Distance covariance in metric spaces. *The Annals of Probability* **41,** 3284–3305 (2013).
9. Gretton, A. & László, G. Consistent nonparametric tests of independence. *Journal of Machine Learning Research* **11,** 1391–1423 (2010).
10. Gretton, A., Herbrich, R., Smola, A., Bousquet, O. & Schölkopf, B. Kernel methods for measuring independence. *Journal of Machine Learning Research* **6,** 2075–2129 (2005).
11. Muandet, K., Fukumizu, K., Sriperumbudur, B. & Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning* **10,** 1–141 (2017).
12. Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* **41,** 2263–2291 (2013).
13. Shen, C. & Vogelstein, J. T. The exact equivalence of distance and kernel methods for hypothesis testing. *arXiv preprint arXiv:1806.05514* (2018).
14. Heller, R., Heller, Y. & Gorfine, M. A consistent multivariate test of association based on ranks of distances. *Biometrika* **100,** 503–510 (2012).

15. Shen, C., Priebe, C. E. & Vogelstein, J. T. From Distance Correlation to Multiscale Graph Correlation. *Journal of the American Statistical Association* **115,** 280–291 (2020).

16. Lee, Y., Shen, C., Priebe, C. E. & Vogelstein, J. T. Network dependence testing via diffusion maps and distance-based correlations. *Biometrika* **106,** 857–873 (2019).

17. Vogelstein, J. T., Bridgeford, E. W., Wang, Q., Priebe, C. E., Maggioni, M. & Shen, C. Discovering and deciphering relationships across disparate data modalities. *eLife* **8,** e41690 (2019).

18. Collingridge, D. S. A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research* **7,** 81–97 (2013).

19. Dwass, M. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics,* 181–187 (1957).

20. Good, P. I. *Permutation, parametric, and bootstrap tests of hypotheses* (Springer Science & Business Media, 2006).

21. Mehta, R., Shen, C., Xu, T. & Vogelstein, J. T. A Consistent Independence Test for Multivariate Time-Series. *arXiv preprint arXiv:1908.06486* (2019).

22. Student. The probable error of a mean. *Biometrika,* 1–25 (1908).

23. Székely, G. J. & Rizzo, M. L. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference* **143,** 1249–1272 (2013).

24. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *Journal of Machine Learning Research* **13,** 723–773 (2012).

25. Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* **52,** 399–433 (1919).

26. Bartlett, M. S. Multivariate analysis. *Supplement to the journal of the royal statistical society* **9,** 176–197 (1947).

27. Warne, R. A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Practical Assessment, Research, and Evaluation* **19,** 17 (2014).

28. Stevens, J. Applied multivariate statistics for the social sciences. Lawrence Erlbaum. *Mahwah, NJ,* 510–1 (2002).

29. Heller, R., Heller, Y., Kaufman, S., Brill, B. & Gorfine, M. Consistent distribution-free k-sample and independence tests for univariate random variables. *The Journal of Machine Learning Research* **17,** 978–1031 (2016).

30. Rizzo, M. L., Székely, G. J., *et al.* Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics* **4,** 1034–1055 (2010).

31. Shen, C., Priebe, C. E. & Vogelstein, J. T. The exact equivalence of independence testing and two-sample testing. *arXiv preprint arXiv:1910.08883* (2019).

32. Pearson, K. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58,** 240–242 (1895).

33. Escoufier, Y. Le traitement des variables vectorielles. *Biometrics,* 751–760 (1973).

34. Robert, P. & Escoufier, Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **25,** 257–265 (1976).

35. Hardoon, D. R., Szedmak, S. & Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* **16,** 2639–2664 (2004).

36. Shen, C., Sun, M., Tang, M. & Priebe, C. E. Generalized Canonical Correlation Analysis for Classification. *Journal of Multivariate Analysis* **130,** 310–322 (2014).

37. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30,** 81–93 (1938).

38. Spearman, C. The proof and measurement of association between two things. *American journal of Psychology* **15,** 72–101 (1904).

39. Agresti, A. *Analysis of ordinal categorical data* (John Wiley & Sons, 2010).

40. Székely, G. J., Rizzo, M. L., *et al.* Partial distance correlation with methods for dissimilarities. *The Annals of Statistics* **42,** 2382–2412 (2014).

41. Hotelling, H. in *Breakthroughs in statistics* 54–65 (Springer, 1992).

42. Micceri, T. The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin* **105,** 156 (1989).

43. Stigler, S. M. Do robust estimators work with real data? *The Annals of Statistics,* 1055–1098 (1977).

44. Carey, G. Multivariate analysis of variance (MANOVA): I. Theory. *Retrieved May* **14,** 2011 (1998).

45. Rencher, A. Methods of multivariate analysis. *DOI* **10,** 66 (2002).

46. Bartlett, M. S. *A note on tests of significance in multivariate analysis* in *Mathematical Proceedings of the Cambridge Philosophical Society* **35** (1939), 180–185.

47. Rao, C. R. Tests of significance in multivariate analysis. *Biometrika* **35,** 58–79 (1948).

48. Garson, G. D. Multivariate glm, manova, and mancova. *Statnotes: Topics in multivariate analysis* (2009).

49. Olson, C. L. On choosing a test statistic in multivariate analysis of variance. *Psychological bulletin* **83,** 579 (1976).

50. Olson, C. L. *A Monte Carlo investigation of the robustness of multivariate analysis of variance* PhD thesis (Thesis (Ph. D.)–University of Toronto, 1973).

51. Fisher, R. A. in *Breakthroughs in statistics* 66–70 (Springer, 1992).

52. Mehta, C. R. & Patel, N. R. Exact inference for categorical data. *Encyclopedia of biostatistics* **2,** 1411–1422 (1998).

53. Good, P. I. Permutation, parametric and bootstrap tests of hypotheses: a practical guide to resampling methods for testing hypotheses. *Permutation, parametric and bootstrap tests of hypotheses: a practical guide to resampling methods for testing hypotheses* **100** (2005).

54. Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. & Sabeti, P. C. Detecting novel associations in large data sets. *science* **334,** 1518–1524 (2011).

55. Gorfine, M., Heller, R. & Heller, Y. Comment on detecting novel associations in large data sets. *Unpublished (available at http://emotion. technion. ac. il/˜ gorfinm/filesscience6. pdf on 11 Nov. 2012)* (2012).

56. Rizzo, M. & Szekely, G. *energy: E-Statistics: Multivariate Inference via the Energy of Data* (2018). https://CRAN.R-project.org/package=energy.

57. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* **11,** 1–20. http://www.jstatsoft.org/v11/i09/ (2004).

58. Brill, B. & Kaufman, S. *HHG: Heller-Heller-Gorfine Tests of Independence and Equality of Distributions* (2019). https://CRAN.R-project.org/package=HHG.

59. Panda, S., Palaniappan, S., Xiong, J., Bridgeford, E. W., Mehta, R., Shen, C. & Vogelstein, J. T. *hyppo: A Comprehensive Multivariate Hypothesis Testing Python Package* 2019. arXiv: 1907.02088 [stat.CO].

A Simulations

Independence test simulations were generated utilizing the following equations.

1. $\texttt{Linear}(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}(-1, 1)^p,$$

$$Y = w^\mathsf{T} X + \kappa\epsilon.$$

2. $\texttt{Exponential}(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}(0, 3)^p,$$

$$Y = \exp\left(w^\mathsf{T} X\right) + 10\kappa\epsilon.$$

3. $\texttt{Cubic}(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}(-1, 1)^p,$$

$$Y = 128\left(w^\mathsf{T} X - \frac{1}{3}\right)^3 + 48\left(w^\mathsf{T} X - \frac{1}{3}\right)^2 - 12\left(w^\mathsf{T} X - \frac{1}{3}\right) + 80\kappa\epsilon.$$

4. $\texttt{Joint Normal}(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Let $\rho = 1/2p$, I_p be the identity matrix of size $p \times p$, J_p be the matrix of ones of size $p \times p$, and $\Sigma = \begin{bmatrix} I_p & \rho J_p \\ \rho J_p & (1 + 0.5\kappa) I_p \end{bmatrix}$. Then,

$$(X, Y) \sim \mathcal{N}(0, \Sigma).$$

5. $\texttt{Step Function}(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}(-1, 1)^p,$$

$$Y = \mathbb{I}\left(w^{\mathsf{T}} X > 0\right) + \epsilon,$$

where $\mathbb{I}$ is the indicator function; that is, $\mathbb{I}(z)$ is unity whenever z is true, and 0 otherwise.

6. `Quadratic`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}(-1, 1)^p,$$

$$Y = \left(w^{\mathsf{T}} X\right)^2 + 0.5\kappa\epsilon.$$

7. `W-Shape`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{U}(-1, 1)^p$,

$$X \sim \mathcal{U}(-1, 1)^p,$$

$$Y = 4\left[\left(\left(w^{\mathsf{T}} X\right)^2 - \frac{1}{2}\right)^2 + \frac{w^{\mathsf{T}} U}{500}\right] + 0.5\kappa\epsilon.$$

8. `Spiral`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{U}(0, 5)$, $\epsilon \sim \mathcal{N}(0, 1)$,

$$X_{|d|} = U \sin(\pi U) \cos^d(\pi U) \text{ for } d = 1, ..., p-1,$$

$$X_{|p|} = U \cos^p(\pi U),$$

$$Y = U \sin(\pi U) + 0.4p\epsilon.$$

9. `Uncorrelated Bernoulli`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{B}(0.5)$, $\epsilon_1 \sim \mathcal{N}\left(0, I_p\right)$, $\epsilon_2 \sim \mathcal{N}(0, 1)$,

$$X \sim \mathcal{B}(0.5)^p + 0.5\epsilon_1,$$

$$Y = (2U - 1) w^{\mathsf{T}} X + 0.5\epsilon_2.$$

10. `Logarithmic`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: For $\epsilon \sim \mathcal{N}\left(0, I_p\right)$,

$$X \sim \mathcal{N}\left(0, I_p\right),$$

$$Y_{|d|} = 2\log_2\left(\left|X_{|d|}\right|\right) + 3\kappa\epsilon_{|d|} \text{ for } d = 1, ..., p.$$

11. `Fourth Root`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}\left(-1, 1\right)^p,$$

$$Y = \left|w^\mathsf{T} X\right|^{1/4} + \frac{\kappa}{4}\epsilon.$$

12. `Sine Period` $4\pi(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: For $U \sim \mathcal{U}\left(-1, 1\right)$, $V \sim \mathcal{N}\left(0, 1\right)^p$, $\theta = 4\pi$,

$$X_{|d|} = U + 0.02pV_{|d|} \text{ for } d = 1, ..., p,$$

$$Y = \sin(\theta X) + \kappa\epsilon.$$

13. `Sine Period` $16\pi(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Same as above except $\theta = 16\pi$ and the noise on Y is changed to $0.5\kappa\epsilon$.

14. `Square`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: For $U \sim \mathcal{U}\left(-1, 1\right)$, $V \sim \mathcal{U}\left(-1, 1\right)$, $\epsilon \sim \mathcal{N}\left(0, 1\right)^p$, $\theta = -\frac{\pi}{8}$,

$$X_{|d|} = U\cos\left(\theta\right) + V\sin\left(\theta\right) + 0.05p\epsilon_{|d|},$$

$$Y_{|d|} = -U\sin\left(\theta\right) + V\cos\left(\theta\right).$$

15. `Diamond`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Same as above except $\theta = \pi/4$.

16. `Two Parabolas`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $\epsilon \sim \mathcal{U}(0,1)$, $U \sim \mathcal{B}(0.5)$,

$$X \sim \mathcal{U}(-1,1)^p,$$

$$Y = \left(\left(w^\mathsf{T} X\right)^2 + 2\kappa\epsilon\right) \cdot \left(U - \frac{1}{2}\right).$$

17. `Circle`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{U}(-1,1)^p$, $\epsilon \sim \mathcal{N}\left(0, I_p\right)$, $r = 1$,

$$X_{|d|} = r\left(\sin\left(\pi U_{|d+1|}\right)\prod_{j=1}^{d}\cos\left(\pi U_{|j|}\right) + 0.4\epsilon_{|d|}\right) \text{ for } d = 1, ..., p-1,$$

$$X_{|d|} = r\left(\prod_{j=1}^{p}\cos\left(\pi U_{|j|}\right) + 0.4\epsilon_{|p|}\right),$$

$$Y_{|d|} = \sin\left(\pi U_{|1|}\right).$$

18. `Ellipse`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Same as above except $r = 5$.

19. `Multiplicative Noise`$(x, y) \in \mathbb{R}^p \times \mathbb{R}^p$: $u \sim \mathcal{N}\left(0, I_p\right)$,

$$x \sim \mathcal{N}\left(0, I_p\right),$$

$$y_{|d|} = u_{|d|} x_{|d|} \text{ for } d = 1, ..., p.$$

20. `Multimodal Independence`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{N}\left(0, I_p\right)$, $V \sim \mathcal{N}\left(0, I_p\right)$, $U' \sim \mathcal{B}(0.5)^p$, $V' \sim \mathcal{B}(0.5)^p$,

$$X = U/3 + 2U' - 1,$$

$$Y = V/3 + 2V' - 1.$$

These have been plotted previously. We can generate 2 sample simulations using the following process:

We do two-sample testing between Z and Z', generated as follows: let $Z = [X|Y]$ be the respective random variables from the independence simulation setup. Then define Q_θ as a rotation matrix for a given angle θ, i.e.,

$$Q_\theta = \begin{bmatrix} \cos\theta & 0 & \dots & -\sin\theta \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sin\theta & 0 & \dots & \cos\theta \end{bmatrix}$$

Then we let

$$Z' = Q_\theta Z^\mathsf{T}$$

be the rotated versions of Z.

This is plotted in the figure below:

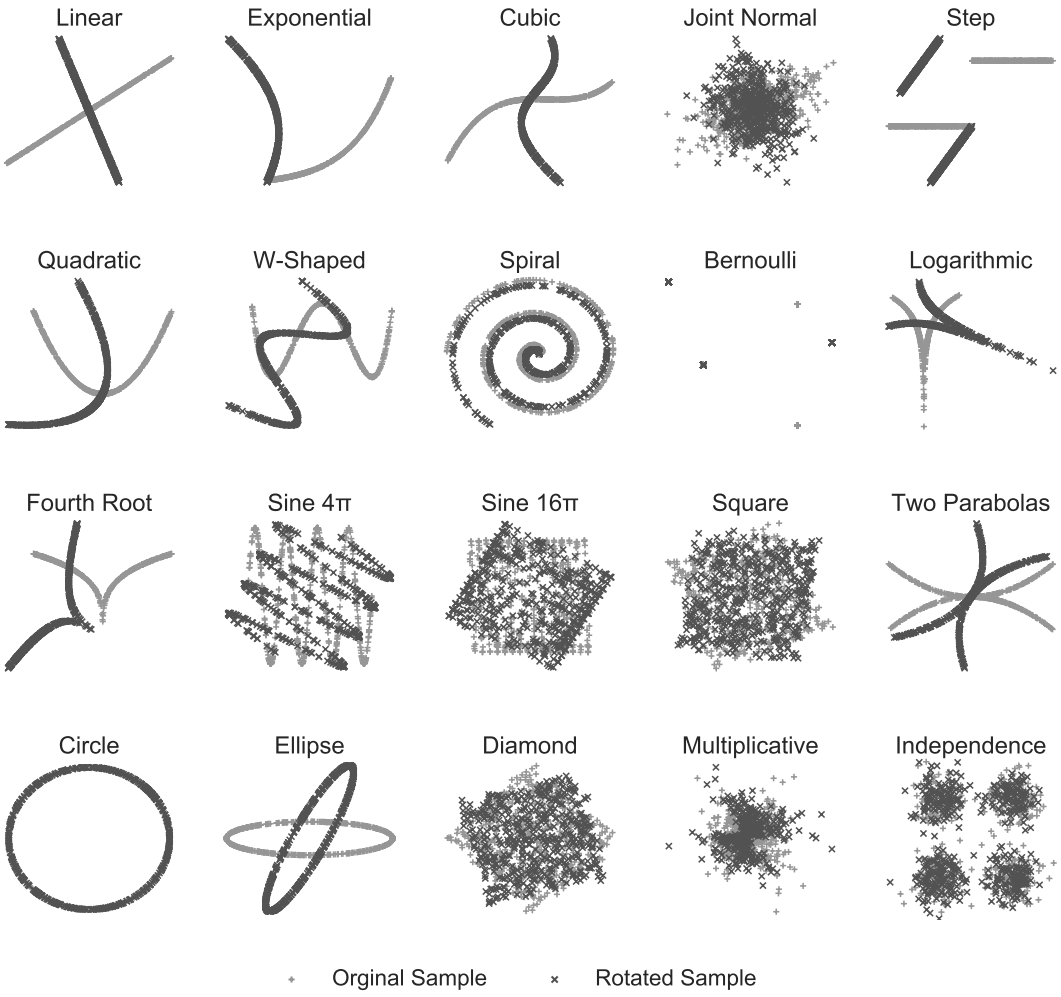

Figure 8: Simulation settings for two-sample power curves. The first dataset (black dots) is 500 samples from each of the 20 different noise-free settings from the hyppo package, the second dataset is the first dataset rotated by 60 degrees. Note that circle simulation has rotational symmetry so the rotation is not evident in 2 dimensions.

Biographical Note

I was born in 1996 in the United States, and have completed my Bachelor's degree in Biomedical Engineering at NC State, my Master's degree in Biomedical Engineering at Johns Hopkins University, and will pursue my Ph.D. in Biomedical Engineering at Johns Hopkins University this fall. I have always had a passion for neuroscience, and have tried to pursue related opportunities my entire career. My goal is to be an active researcher in Biomedical data science and to understand problems from a theoretical perspective as well as an applied one.

Throughout my undergraduate career I worked with Dr. Leslie Sombers at NC State where I grew a passion for research and learned a lot of speaking skills having the honor of presenting at many international research conferences. The pinnacle of this experience was my first publication, as a second author, which discussed a new electro-chemical sensor.

These past 2 years, I have worked extensively with Dr. Joshua T. Vogelstein in the Department of Biomedical Engineering. My projects included those listed in these thesis as well as another on a random forest based independence test. As I pursue a Ph.D., I hope to enter the exciting and fast-growing field of machine learning research, and continuing learning for the rest of my life.

Sambit Panda