`hyppo`: A Comprehensive Multivariate Hypothesis Testing Python Package

Sambit Panda[1], Satish Palaniappan[1], Junhao Xiong[1], Eric W. Bridgeford[2], Ronak Mehta[1], Cencheng Shen[3], and Joshua T. Vogelstein[1] *

Abstract. We introduce `hyppo`, a unified library for performing multivariate hypothesis testing, including independence, two-sample, and *k*-sample testing. While many multivariate independence tests have `R` packages available, the interfaces are inconsistent and most are not available in `Python`. `hyppo` includes many state of the art multivariate testing procedures. The package is easy-to-use and is flexible enough to enable future extensions. The documentation and all releases are available at https://hyppo.neurodata.io.

Key words. Python, multivariate, independence, *k*-sample, hypothesis

1 Introduction Examining and identifying relationships between sets of high-dimensional variables is critical to advance understanding and planning of future numerical and physical experiments. Hypothesis testing enables formally testing models to identify such discrepancies.

Many correlation measures have been proposed the problem of independence testing, such as Pearson's correlation [1], but many are unsuited to detect nonlinear and high-dimensional dependence structures within data. Recently, several statistics have been proposed that operate well on high-dimensional (potentially non-Euclidean) data, such as distance correlation [2–5] and Hilbert-Schmidt independence criterion [6–8], which are actually exactly equivalent in Sejdinovic et al. [9], Shen and Vogelstein [10]. Heller, Heller and Gofrine proposed another nonparametric independence test with particularly high power in certain nonlinear relationships [11]. 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 [12, 13], 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 [14]. For each of these tests, p-values can be calculated using a random permutation test [15–17]. These tests can be modified and extended to such applications as time-series testing [18].

To approach the problem of two-sample testing, Student's t-test [19] is traditionally used, while a few nonparametric alternatives have been proposed that operate well on multivariate, nonlinear data such as Energy [20], and maxmimal mean discrepency [21], and Heller Heller and Gorfine's test [11]. The two-sample testing problem can be generalized to the *k*-sample testing problem and here analysis of variance (ANOVA) [22] or its multivariate analogue multivariate ANOVA (MANOVA) [23] can be used, but these statistics either fail to, or operate poorly upon, multivariate and nonlinear data. In addition, ANOVA and MANOVA in particular suffer from fundamental assumptions that are not generally present in real data [24, 25]. There are a few nonparametric alternatives to ANOVA and MANOVA, such as multivariate *k*-sample Heller Heller Gorfine [26], and distance components (DISCO) [27]. Recently, Shen et al. [28] 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 manuscript 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` also provides benchmarks for each of these tests by comparing statistical power over many statistical models. The contribution of this manuscript is therefore to provide: (1) an overview of the library and examples of how to use some of the tests in the package, and (2) comparisons of the test statistics and wall times with similar `R` packages.

*Corresponding author: jovo@jhu.edu [1] Johns Hopkins University, Baltimore, MD [2] Johns Hopkins School of Public Health, Baltimore, MD [3] University of Delaware, Newark, DE

2 Library Overview

Inspired by the desire to allow for convenient use of these independence tests, hyppo has been developed as a comprehensive independence and k-sample testing package. The package structured is modeled on the scikit-learn and energy R packages' API. Links to source code, documentation, and tutorials can be found here: https://hyppo.neurodata.io.

Included Tests To make hyppo a more comprehensive independence testing package, some existing methods have been implemented from established tests or are wrappers of existing implementations. Shen and Vogelstein [10] has shown that distance and kernel methods are equivalent and thus, we have 1 implementation that is able to perform both with a proper bijective transformation. We have implemented k-sample tests as specified in Shen et al. [28] and every algorithm in the following list except those listed in the first bullet can be used as a k-sample test this way. The included algorithms are:

- Univariate classical independence tests—these are wrappers of existing methods in scipy—Pearson correlation (PEARSON) [1], Spearman (SPEARMAN) [29], Kendall (KENDALL) [30].
- Multivariate generalizations of Pearson's product moment correlation: RV [31, 32] and Cannonical correlation analysis (CCA) [33].
- Heller-Heller-Gorfine (HHG) [11]: Multivariate distance-based test.
- Distance correlation (DCORR), both biased [4] and unbiased [34].
- Hilbert-Schmidt independence criterion (HSIC), both biased and unbiased [35] kernel-based statistics.
- Maximum mean discrepancy (MMD) [21]: A kernel two-sample test.
- ENERGY [20]: A distance two-sample test.
- Distance components (DISCO) [27]: A distance-based k-sample test.
- Multiscale graph correlation (MGC) [14]: An independence tests that combines k-nearest neighbors and energy statistics. Recently, MGC has been accepted into scipy.stats and this implementation wraps the scipy implementation.

A number of algorithms have been implemented that lack an open source implementation elsewhere. These include:

- Kernel k-sample tests: k-sample HSIC and k-sample MGC formulated using the k-sample formulation as above [28].
- Fast Implementations of DCORR (FAST DCORR) [36]: A chi-square approximation to DCORR and MGC when calculating the test statistic.
- Time-series MGC and DCORR: Applying MGC and DCORR to time-series data.

Structure of hyppo The modules of hyppo are: independence, ksample, time_series, and sims. 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.

3 Benchmarks

Wall Time Comparisons Figure 1a shows the computational efficiency of hyppo's implementations against existing implementations in commonly used R packages—specifically energy [37], kernlab [38], and HHG [39]. 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 [10]) 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.

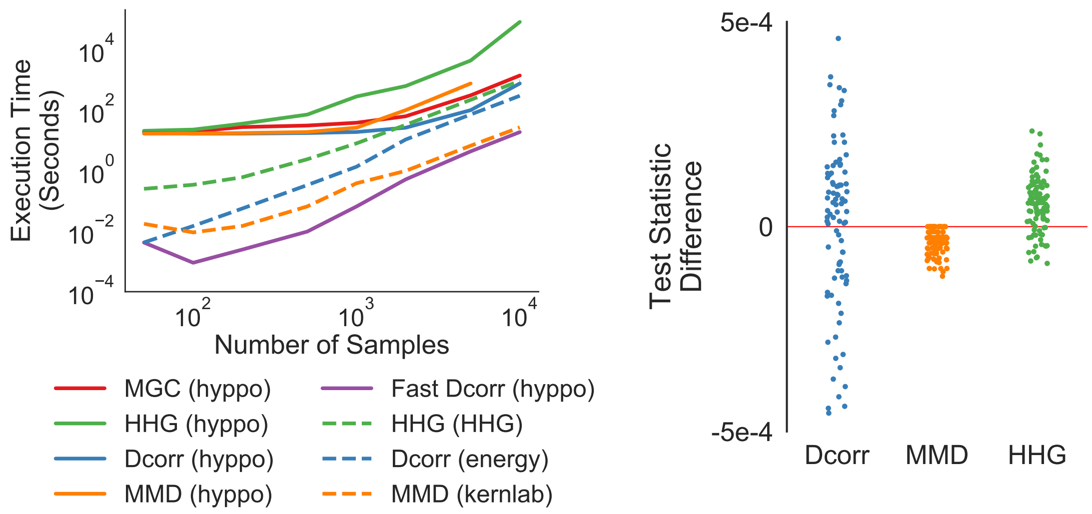

Figure 1: Benchmarks of `hyppo` implementations against corresponding `R` implementations. Average wall times (over 3 repetitions) (left) are shown for Dcorr in `energy` and `kernlab` as compared against `hyppo` implementations of Mgc, Dcorr, Fast Mgc, and Fast Dcorr. Test statistic comparisons (right)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

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 [37]. Mmd from the `kernlab` package [38], and Hhg from the `HHG` package [39]. 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. Test statistics are nearly equivalent for each implementation.

4 Conclusion `hyppo` is an extensive and extensible open-source `Python` package for multivariate hypothesis testing. As `hyppo` continues to grow and add functionality, it will enhance tools scientists use when determining relationships within their investigations.

Acknowledgements This work is graciously supported by the Defense Advanced Research Projects Agency (DARPA) Lifelong Learning Machines program through contract FA8650-18-2-7834, and the National Science Foundation award DMS-1921310. The authors would like to acknowledge the NeuroData Design class and the NeuroData lab at Johns Hopkins University for helpful feedback throughout the development process.

References

[1] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

[2] Gábor J Székely and Maria L Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

[3] Gábor J Székely and Maria L Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013.

[4] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[5] Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

[6] Arthur Gretton and Györfi László. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11(Apr):1391–1423, 2010.

[7] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005.

[8] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[9] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

[10] Cencheng Shen and Joshua T Vogelstein. The exact equivalence of distance and kernel methods for hypothesis testing. *arXiv preprint arXiv:1806.05514*, 2018.

[11] Ruth Heller, Yair Heller, and Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2012.

[12] Cencheng Shen, Carey E Priebe, and Joshua T Vogelstein. From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association*, pages 1–22, 2019.

[13] Youjin Lee, Cencheng Shen, Carey E Priebe, and Joshua T Vogelstein. Network dependence testing via diffusion maps and distance-based correlations. *Biometrika*, 106(4):857–873, 2019.

[14] Joshua T Vogelstein, Eric W Bridgeford, Qing Wang, Carey E Priebe, Mauro Maggioni, and Cencheng Shen. Discovering and deciphering relationships across disparate data modalities. *eLife*, 8:e41690, 2019.

[15] Dave S Collingridge. A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research*, 7(1):81–97, 2013.

[16] Meyer Dwass. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pages 181–187, 1957.

[17] Phillip I Good. *Permutation, parametric, and bootstrap tests of hypotheses*. Springer Science & Business Media, 2006.

[18] Ronak Mehta, Cencheng Shen, Ting Xu, and Joghua T Vogelstein. A consistent independence test for multivariate time-series. *arXiv preprint arXiv:1908.06486*, 2019.

[19] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[20] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[22] Ronald A Fisher. Xv.âĂŤthe correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2): 399–433, 1919.

[23] Maurice S Bartlett. Multivariate analysis. *Supplement to the journal of the royal statistical society*,

9(2):176–197, 1947.

[24] Russell Warne. A primer on multivariate analysis of variance (manova) for behavioral scientists. *Practical Assessment, Research, and Evaluation*, 19(1):17, 2014.

[25] JP Stevens. Applied multivariate statistics for the social sciences. lawrence erlbaum. *Mahwah, NJ*, pages 510–1, 2002.

[26] Ruth Heller, Yair Heller, Shachar Kaufman, Barak Brill, and Malka Gorfine. Consistent distribution-free k-sample and independence tests for univariate random variables. *The Journal of Machine Learning Research*, 17(1):978–1031, 2016.

[27] Maria L Rizzo, Gábor J Székely, et al. Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010.

[28] Cencheng Shen, Carey E Priebe, and Joshua T Vogelstein. The exact equivalence of independence testing and two-sample testing. *arXiv preprint arXiv:1910.08883*, 2019.

[29] Charles Spearman. The proof and measurement of association between two things. *American journal of Psychology*, 15(1):72–101, 1904.

[30] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[31] Yves Escoufier. Le traitement des variables vectorielles. *Biometrics*, pages 751–760, 1973.

[32] Paul Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(3):257–265, 1976.

[33] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[34] Gábor J Székely, Maria L Rizzo, et al. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

[35] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

[36] Cencheng Shen. High-dimensional independence testing and maximum marginal correlation. *arXiv preprint arXiv:2001.01095*, 2020.

[37] Maria Rizzo and Gabor Szekely. *energy: E-Statistics: Multivariate Inference via the Energy of Data*, 2018. URL https://CRAN.R-project.org/package=energy. R package version 1.7-5.

[38] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. URL http://www.jstatsoft.org/v11/i09/.

[39] Barak Brill and Shachar Kaufman. *HHG: Heller-Heller-Gorfine Tests of Independence and Equality of Distributions*, 2019. URL https://CRAN.R-project.org/package=HHG. R package version 2.3.2.