Statistica Sinica **36** (2026), 1-24
doi:https://doi.org/10.5705/ss.202023.0408

A ROBUST FRAMEWORK FOR GRAPH-BASED TWO-SAMPLE TESTS USING WEIGHTS

Yichuan Bai* and Lynna Chu

Iowa State University

Abstract: Graph-based tests are a class of non-parametric two-sample tests useful for analyzing high-dimensional data. The test statistics are constructed from similarity graphs (such as K-minimum spanning tree), and consequently, their performance is sensitive to the structure of the graph. When the graph has problematic structures (for example, hubs), as is common for high-dimensional data, this can result in low power and unstable performance among existing graph-based tests. We address this challenge by proposing new test statistics that are robust to problematic structures of the graph and can provide reliable inferences. We employ an edge-weighting strategy using intrinsic characteristics of the graph that are computationally simple and efficient to obtain. The limiting null distribution of the robust test statistics is derived and shown to work well for finite sample sizes. Simulation studies and data analysis of Chicago taxi-trip travel patterns demonstrate the new tests' improved performance across a range of settings.

Key words and phrases: Curse of dimensionality, graph-based tests, high-dimensional data, non-parametric tests, robustness, similarity graphs

1. Introduction

We focus on testing the equality of distributions for observations in the high dimensional setting, where the dimension of the observation d may be much larger than the sample size N. Suppose we have two samples $\{\mathbf{X_1}, \ldots, \mathbf{X_{n_1}}\}$ and $\{\mathbf{Y_1}, \ldots, \mathbf{Y_{n_2}}\}$ of d-dimensional observations that are independently and identically distributed from unknown distributions F_X and F_Y, respectively. The two-sample problem aims to test $H_0 : F_X = F_Y$ against an omnibus alternative $H_1 : F_X \neq F_Y$. This is a classic statistical problem but made more challenging by the increasing complexity of modern data, where observations can be high-dimensional data objects ($d >> N$). In this setting, it is often intractable to express or estimate F_X and F_Y directly due to the curse of dimensionality. Substantial developments have been made by the contemporary statistics community to address such challenges. For example, non-parametric two-sample tests for multivariate and high-dimensional data have been proposed using distances (Baringhaus and Franz, 2004; Székely and Rizzo, 2004; Biswas and Ghosh, 2014; Li, 2018), generalized ranking (Liu and Singh, 1993; Hall and Tajvidi, 2002), and

*Corresponding author. E-mail: ycbai@alumni.iastate.edu