Statistica Sinica **35** (2025), 1-22
doi:https://doi.org/10.5705/ss.202023.0069

HIGH DIMENSIONAL BEHAVIOUR OF SOME TWO-SAMPLE TESTS BASED ON BALL DIVERGENCE

Bilol Banerjee and Anil K. Ghosh*

Indian Statistical Institute, Kolkata

Abstract: We propose some two-sample tests based on ball divergence and investigate their high dimensional behaviour. First, we consider the High Dimension, Low Sample Size (HDLSS) setup. Under appropriate regularity conditions, we establish the consistency of these tests in the HDLSS regime, where the dimension grows to infinity while the sample sizes from the two distributions remain fixed. Next, we show that these conditions can be relaxed when the sample sizes also increase with the dimension, and in such cases, consistency can be proved even for shrinking alternatives. We use a simple example to show that even when there are no consistent tests in the HDLSS regime, the proposed tests can be consistent if the sample sizes increase with the dimension at an appropriate rate. This rate is obtained by establishing the minimax rate optimality of these tests over a certain class of alternatives. Several simulated and benchmark data sets are analyzed to compare the empirical performance of these tests with some state-of-the-art methods available for testing the equality of two high dimensional distributions.

Key words and phrases: Ball divergence, energy statistics, high dimensional asymptotics, minimax rate optimality, permutation tests, shrinking alternatives.

1. Introduction

In a two-sample problem, we test for the equality of two d-dimensional distributions F and G based on n independent copies $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X} \sim F$ and m independent copies $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ of $\mathbf{Y} \sim G$. This problem is well investigated in the literature, and several tests are available for it. In the parametric regime, we often assume F and G to be Gaussian and test for the equality of their location and/or scale parameters. Several nonparametric tests are also available, especially for $d = 1$. While the Wilcoxon-Mann-Whitney test is used for the univariate two-sample location problem, the Wald-Wolfowitz run test, the Kolmogorov-Smirnov (KS) test, and the Camer-von-Mises (CVM) test (Hollander, Wolfe and Chicken, 2014; Gibbons and Chakraborti, 2011) are applicable to general two-sample problems.

Using the idea of a minimum spanning tree, Friedman and Rafsky (1979) generalized the run test and the KS test to higher dimensions. Baringhaus and Franz (2004) proposed a test based on inter-point distances, which can

*Corresponding author. E-mail: akghosh@isical.ac.in