ON SIMULTANEOUS CALIBRATION OF TWO-SAMPLE $t$-TESTS FOR HIGH-DIMENSION LOW-SAMPLE-SIZE DATA

Chunming Zhang, Shengji Jia and Yongfeng Wu

University of Wisconsin, Madison

Abstract: The exact distribution is typically unavailable for a two-sample $t$-statistic in a single test for equal population means if we have non-Gaussian samples, unequal population variances, or unequal sample sizes $n_1$ and $n_2$. In this case, a calibration method using a reference distribution offers a practically feasible substitute. This study simultaneously calibrates a diverging number $m$ of two-sample $t$-statistics for inferences of significance in high-dimensional data from a small sample. For the Gaussian calibration method, we demonstrate the following. First, the simultaneous “general” two-sample $t$-statistics achieve the overall significance level, as long as $\log(m)$ increases at a strictly slower rate than $(n_1 + n_2)^{1/3}$ as $n_1 + n_2$ diverges. Second, directly applying the same calibration method to simultaneous “pooled” two-sample $t$-statistics may substantially lose the overall level accuracy. The proposed “adaptively pooled” two-sample $t$-statistics overcome such incoherence, while operating as simply and performing as well as the “general” two-sample $t$-statistics. Third, we propose a “two-stage” $t$-test procedure to effectively alleviate the skewness commonly encountered in various two-sample $t$-statistics in practice, thus increasing the calibration accuracy. Lastly, we discuss the implications of these results using simulation studies and real-data applications.

Key words and phrases: Familywise error rate, multiple hypothesis testing, overall significance level, simultaneous inference, skewness.

1. Introduction

With the advancement of high-throughput technology, large-scale simultaneous inference procedures [Bourgon, Gentleman and Huber (2010); Efron (2010); Liang and Nettleton (2012); Leek and Storey (2008); Zhang, Fan and Yu (2011); Zhao, Wang and Wei (2013)] arise naturally from high-dimensional data from small samples, with wide applications in biology, genetics, astronomy, economics, and neuroscience research among others. This problem is characterized by simultaneously carrying out a large number of hypothesis tests, where each test

Corresponding author: Chunming Zhang, Department of Statistics, University of Wisconsin, Madison, 53706, USA. E-mail: cmzhang@stat.wisc.edu.