Statistica Sinica **33** (2023), 1-24
doi:https://doi.org/10.5705/ss.202020.0405

NEW TESTS FOR HIGH-DIMENSIONAL LINEAR REGRESSION BASED ON RANDOM PROJECTION

Changyu Liu, Xingqiu Zhao and Jian Huang

The Hong Kong Polytechnic University

Abstract: We consider the problem of detecting significance in high-dimensional linear models, in which the dimension of the regression coefficient is greater than the sample size. We propose novel test statistics for hypothesis tests of the global significance of the linear model, as well as for the significance of part of the regression coefficients. The new tests are based on randomly projecting the high-dimensional data onto a low-dimensional space, and then working with the classical F-test using the projected data. An appealing feature of the proposed tests is that they have a simple form and are computationally easy to implement. We derive the asymptotic local power functions of the proposed tests and compare them with the existing methods for hypothesis testing in high-dimensional linear models. We also provide a sufficient condition under which our proposed tests have higher asymptotic relative efficiency. Simulation studies evaluate the finite-sample performance of the proposed tests and demonstrate that it outperforms existing tests in the models considered. Lastly, we illustrate the proposed tests by applying them to real high-dimensional gene expression data.

Key words and phrases: High-dimensional inference, hypothesis testing, linear model, random projection, relative efficiency.

1. Introduction

High-dimensional data are now routinely encountered in many fields of scientific research. For example, in genomic studies, the dimension of data such as gene expression and genetic marker data is typically far greater than the sample size. A common feature of high-dimensional data is that the data dimension p can be greater than the sample size n. This phenomenon brings challenges to classical statistical analysis, even in many basic settings. For example, the Hotelling T^2 statistic for the two-sample testing problem is not well defined when p is larger than n, because the sample covariance matrix is no longer invertible in this setting. In high-dimensional linear regression models, existing methods for statistical inference about regression coefficients are no longer applicable. There-

Corresponding author: Xingqiu Zhao, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. E-mail: xingqiu.zhao@polyu.edu.hk.