Statistica Sinica **36** (2026), 1-20
doi:https://doi.org/10.5705/ss.202021.0406

LARGE-SCALE MULTIPLE TESTING FOR MATRIX-VALUED DATA UNDER CROSS-DEPENDENCY

Shiyu Zhang, Xu Han and Sanat K. Sarkar*

Temple University

Abstract: High-dimensional inference based on matrix-valued data has drawn increasing attention in modern statistical research, yet not much progress has been made in large-scale multiple testing specifically designed for analyzing such data sets. Motivated by this, we consider in this article an electroencephalography (EEG) experiment that produces matrix-valued data and presents a scope of developing novel matrix-valued data based multiple testing methods that are of importance in such an experiment. The row-column cross-dependency of observations appearing in a matrix form, referred to as cross-dependency, is one of the main challenges in the development of such methods. We address this challenge by assuming matrix normal distribution for the observations at each of the independent matrix data-points. This allows us to capture the underlying cross-dependency informed through the row- and column-covariance matrices and develop methods that are potentially better than the corresponding one obtained by vectorizing each data point and thus ignoring the cross-dependency. Given a fixed thresholding procedure with unknown cross covariance matrices, we consider approximating the false discovery proportion capturing the underlying cross-dependency with statistical accuracy and propose two methods of doing so. While one of these methods is a general approach under cross-dependency, the other one provides more computational efficiency for higher dimensionality. Extensive numerical studies illustrate the superior performance of the proposed methods over the principal factor approximation method of Fan and Han (2017). The proposed methods have been further applied to the aforementioned EEG data.

Key words and phrases: Cross dependency, electroencephalogram, false discovery proportion, large-scale multiple testing, matrix-valued data.

1. Introduction

Large-scale multiple testing is an integral part of statistical investigations in the modern era of Big Data-driven scientific research with statisticians/data scientists frequently encountering simultaneous testing of tens of thousands or even hundreds of thousands of hypotheses in such research. Despite substantial growth of research in multiple testing over the past few decades (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Sarkar, 2002; Storey, 2002; Efron,

*Corresponding author. E-mail: sanat@temple.edu