Statistica Sinica **34** (2024), 1-23
doi:https://doi.org/10.5705/ss.202022.0069

ROBUST RANK CANONICAL CORRELATION ANALYSIS FOR MULTIVARIATE SURVIVAL DATA

Di He, Yong Zhou and Hui Zou*

Nanjing University, East China Normal University and University of Minnesota

Abstract: Canonical correlation analysis (CCA) is widely applied in statistical analysis of multivariate data to find associations between two sets of multidimensional variables. However, we often cannot use CCA directly for survival data or their monotone transformations, owing to right-censoring in the data. In this paper, we propose a new robust rank CCA (RRCCA) method based on Kendall's τ correlation, and adjust it to deal with multivariate survival data, without requiring any model assumptions. Owing to the nature of rank correlation, the RRCCA is invariant against monotone transformations of the data. We establish the estimation consistency of the RRCCA approach under weak conditions. Simulation studies demonstrate the superior performance of the RRCCA in terms of estimation accuracy and empirical power. Lastly, we demonstrate the proposed method by applying it to Stanford heart transplant data.

Key words and phrases: Canonical correlation analysis, inverse probability of censoring weighting, Kendall's τ correlation, right-censoring.

1. Introduction

Canonical correlation analysis (CCA), introduced by Hotelling (1936), is a well-known statistical technique for finding associations between two sets of multidimensional variables. It searches for linear projections of each set of variables, such that the projected variables are maximally correlated. Extensions of the classical CCA have been proposed for particular kinds of practical data sets. For example, Akaho (2001) developed a kernel CCA for discovering nonlinear correlations among variables, Vía, Santamaría and Pérez (2007) proposed a generalization of the CCA that can handle several data sets. Sparse CCAs (Witten, Tibshirani and Hastie (2009); Hardoon and Shawe-Taylor (2011); Mai and Zhang (2019); Chen et al. (2020)) have been proposed for high-dimensional data sets, and supervised CCAs (Witten and Tibshirani (2009); Golugula et al. (2011)) are used when the two sets of variables are associated with the outcomes.

In medicine, demography, economics, and other fields, available data on the time to some event are not always exact and complete. These survival times,

*Corresponding author.