Statistica Sinica **37** (2027), 1-22
doi:https://doi.org/10.5705/ss.202024.0261

SEMI-SUPERVISED REGRESSION ANALYSIS WITH MODEL MISSPECIFICATION AND HIGH-DIMENSIONAL DATA

Ye Tian[1], Peng Wu[2] and Zhiqiang Tan[*1]

[1]*Rutgers University and* [2]*Beijing Technology and Business University*

Abstract: The accessibility of vast volumes of unlabeled data has sparked growing interest in semi-supervised learning (SSL) and covariate shift transfer learning (CSTL). In this paper, we present an inference framework for estimating regression coefficients in conditional mean models within both SSL and CSTL settings, while allowing for the misspecification of conditional mean models. We develop an augmented inverse probability weighted (AIPW) method, employing regularized calibrated estimators for both propensity score (PS) and outcome regression (OR) nuisance models, with PS and OR models being sequentially dependent. We show that when the PS model is correctly specified, the proposed estimator achieves consistency, asymptotic normality, and valid confidence intervals, even with possible OR model misspecification and high-dimensional data. Moreover, by suppressing detailed technical choices, we demonstrate that previous methods can be unified within our AIPW framework. Our theoretical findings are verified through extensive simulation studies and a real-world data application.

Key words and phrases: Augmented inverse probability weighted estimator, covariate shift transfer learning, high-dimensional data, semi-supervised learning.

1. Introduction

In recent years, vast volumes of unlabeled data have become increasingly accessible, sparking growing interest in how to leverage these data in both academic research and industrial applications. One of the active areas of research is semi-supervised learning (SSL). In addition, covariate shift transfer learning (CSTL) also exploits information from unlabeled data. Both have various application scenarios like computer vision (Sohn et al., 2020; Zhou and Levine, 2021; Zheng et al., 2022), natural language process (Chen and Huang, 2016; Ruder et al., 2019; Zhao et al., 2022), causal inference (Alvari et al., 2019; Aloui et al., 2023; Zhang, Chakrabortty and Bradic, 2023), healthcare data analysis (Castro, Walker and Glocker, 2020; Liu et al., 2023; Tang et al., 2024), etc.

In both SSL and CSTL settings, we have access to a labeled dataset $\mathcal{L}$ and an unlabeled dataset $\mathcal{U}$, where the labeled dataset $\mathcal{L}$ contains observations with both the covariates $\boldsymbol{X}$ and the outcome Y, while the unlabeled dataset $\mathcal{U}$ consists

*Corresponding author. E-mail: ztan@stat.rutgers.edu.