Statistica Sinica **32** (2022), 2023-2046
doi:https://doi.org/10.5705/ss.202020.0251

POISSON REGRESSION WITH ERROR CORRUPTED HIGH DIMENSIONAL FEATURES

Fei Jiang and Yanyuan Ma

The University of California, San Francisco and Pennsylvania State University

Abstract: Features extracted from aggregated data are often contaminated with errors. Errors in these features are usually difficult to handle, especially when the feature dimension is high. We construct an estimator of the feature effects in the context of a Poisson regression with a high dimensional feature and additive measurement errors. The procedure penalizes a target function that is specially designed to handle measurement errors. We perform optimization within a bounded region. Benefiting from the convexity of the constructed target function in this region, we establish the theoretical properties of the new estimator in terms of algorithmic convergence and statistical consistency. The numerical performance is demonstrated using simulation studies. We apply the method to analyze the possible effect of weather on the number of COVID-19 cases.

Key words and phrases: Composite gradient descent, COVID-19, non-convex optimization, Poisson regression, measurement error.

1. Introduction

Measurement errors frequently occur to features extracted from aggregated data sets, such as average temperatures from multiple sensors, owing to the loss of raw data information after the data aggregation. The measurement error issue for count outcome prediction has gained great attention in infectious disease studies where numerous data are collected to predict disease spread. For example, with the recent outbreak of the COVID-19 pandemic, there is some hope that the pandemic will ease when the weather becomes warmer. However, conclusions on the association between climate and COVID-19 infection are varied and controversial. For example, Tosepu et al. (2020) showed that temperature has a positive association with COVID-19 cases, whereas Jüni et al. (2020) showed that there is no significant association between climate and COVID-19 cases. Nevertheless, none of these studies considered the potential error contamination of the climate data. For example, weather components such as temperature and precipitation vary within a county, whereas the COVID-19 cases are usually summarized at

Corresponding author: Fei Jiang, School of Medicine, UCSF, CA 94158, USA. E-mail: fei.jiang@ucsf.edu.