Statistica Sinica **36** (2026), 1-25
doi:https://doi.org/10.5705/ss.202023.0195

DYNAMIC STATISTICAL LEARNING IN MASSIVE DATASTREAMS

Jingshen Wang, Lilun Du*, Changliang Zou and Zhenke Wu

University of California, Berkeley, City University of Hong Kong, Nankai University and University of Michigan

Abstract: Technological advances have necessitated statistical methodologies for analyzing large-scale datastreams comprising multiple indefinitely time series. This article proposes a dynamic tracking and screening (DTS) framework for online learning and model updating. Utilizing the sequential nature of datastreams, a robust estimation approach is developed under a linear varying coefficient model framework. This accommodates unequally-spaced design points and updates coefficient estimates without storing historical data. A data-driven choice of an optimal smoothing parameter is proposed, alongside a new multiple testing procedure for the streaming environment. Statistical guarantees of the procedure are provided, along with simulation studies on its finite-sample performance. The methods are demonstrated through a mobile health example estimating when subjects' sleep and physical activities unusually influence their mood.

Key words and phrases: Consistency, kernel smoothing, multiple testing, varying coefficient.

1. Introduction

1.1. Background and motivation

Highly developed information and sensor technologies constantly generate and store massive longitudinal data sets that become available sequentially at a high frequency. Ranging from telecommunications (Black and Hickey, 2003), environmental monitoring (Guerriero, Willett and Glaz, 2009), retail banking (Tsung, Zhou and Jiang, 2007), health care (Spiegelhalter et al., 2012), and network monitoring (Vaughan, Stoev and Michailidis, 2013), such a type of data collection is pervasive and is referred to as streaming data throughout this article. Other than the high-frequency feature, as massive datastreams are often collected from distinct classes of subjects often in highly dynamic real-life environments, it is commonly believed that they may contain a growing number of irregular patterns (Gama, 2010).

In this context, a statistical methodology that is relevant to streaming data analysis often pertains to algorithms that enable us to

*Corresponding author. E-mail: lilundu@cityu.edu.hk