Statistica Sinica **37** (2027), 1-22
doi:https://doi.org/10.5705/ss.202023.0216

SIMULTANEOUS JUMP DETECTION FOR MULTIPLE SEQUENCES VIA SCREENING AND MULTIPLE TESTING

Shengji Jia and Chunming Zhang*

Shanghai Lixin University of Accounting and Finance and University of Wisconsin-Madison

Abstract: The estimation of nonparametric discontinuous regression function is fundamental in many applied fields, but challenges arise when the number of jumps (or discontinuities) is large and unknown. We propose a new jump detection method, via the consecutive screening and multiple testing (SaMT) algorithm, for simultaneously estimating the unknown number of jump points and detecting their locations in the flexible nonparametric regression model, guaranteeing the desired accuracy. The initial jump candidates are obtained in the *consecutive screening* procedure combined with locally-linear smoothing method. To further assess the significance of an individual jump candidate, we develop a novel test based on profile likelihood inference. The ultimate selection of relevant jump points is conducted in a *multiple testing* procedure, which eliminates irrelevant jump points with large variations, due to heteroscedastic errors, from jump candidates. Moreover, we generalize the SaMT algorithm to detect common jump points shared across multiple aligned sequences. The proposed method is easy to implement, flexible in bandwidth and threshold selection, and outperforms existing approaches in simulations and real-data applications.

Key words and phrases: Copy number variation, false discovery rate, jump regression analysis, locally linear smoothing, Wald statistic.

1. Introduction

In many areas, such as environmental statistics, genetics, finance and engineering, very long and noisy sequences of data will arise. Examples include high-throughput sequencing data in genetics and high-frequency financial data in econometrics. Such data usually have discontinuities or jumps which are important data structures, and the goal is to identify and understand structural variations, e.g., from a normal number to an excessive number of chromosome copies in genetics, or structural changes in trends, e.g., from a bull market to a bear market in finance. Thus, identifying multiple jump points in the regression function is a fundamental and challenging problem.

In the literature, the issue of jump detection was investigated using different approaches, including locally-linear smoothing (Grégoire and Hamrouni, 2002; Xia and Qiu, 2015), splines (Ma and Yang, 2011), and wavelets (Wang,

*Corresponding author. E-mail: cmzhang@stat.wisc.edu