Statistica Sinica **37** (2027), 1-21
doi:https://doi.org/10.5705/ss.202024.0222

DISTRIBUTED INFERENCE FOR TAIL RISKS

Liujun Chen, Deyuan Li, and Chen Zhou*

University of Science and Technology of China,
Fudan University and Erasmus University Rotterdam

Abstract: For measuring tail risk with scarce extreme events, extreme value analysis is often invoked as the statistical tool to extrapolate to the tail of a distribution. The presence of large datasets benefits tail risk analysis by providing more observations for conducting extreme value analysis. However, large datasets can be stored distributedly preventing the possibility of directly analyzing them. In this paper, we introduce a comprehensive set of tools for examining the asymptotic behavior of tail empirical and quantile processes in the setting where data is distributed across multiple sources, for instance, when data are stored on multiple machines. Utilizing these tools, one can establish the oracle property for most distributed estimators in extreme value statistics in a straightforward way. We provide various examples to demonstrate the practicality and value of our proposed toolkit.

Key words and phrases: KMT inequality, oracle property, tail empirical process, tail quantile process.

1. Introduction

Financial and climate risk management requires risk forecasting for rare but high-impact events, typically referred to as extreme events. Extreme value analysis, statistical methods for analyzing the tail of a distribution, is a useful tool for modeling and analyzing such extremes (de Haan and Ferreira, 2006). In this paper, we consider tail risk analysis using a large dataset that is distributedly stored at various locations.

While the availability of large datasets in general benefits statistical analysis, such as extreme value analysis, it also presents at least three practical challenges to implementing conventional statistical procedures. Firstly, a combined dataset might not be available to one end user due to privacy concerns such as analyzing large insurance claims across various insurance companies (Embrechts, Klüppelberg and Mikosch, 2013). Since insurance companies are contracted to protect the privacy of their customers, it is impossible to combine all claims from different insurers into one massive dataset. Secondly, the computation cost of analyzing a massive dataset can be expensive when implementing statistical procedures involving an optimization algorithm, such as maximum likelihood or loss minimization. Thirdly, storage constraints can arise when dealing with

*Corresponding author. E-mail: zhou@ese.eur.nl